SAPIENZA
Università di Roma
Facoltà di Scienze Matematiche Fisiche e Naturali

DOTTORATO DI RICERCA
IN GENETICA E BIOLOGIA MOLECOLARE

XXXIII Ciclo
(A.A. 2019/2020)

ANALYSIS OF NON-ALLELIC GENE CONVERSION IN THE EVOLUTION OF HUMAN MSY PALINDROMES

Dottorando
Maria Bonito
Maria Bonito

Docente guida
Prof. Fulvio Cruciani
Fulvio Cruciani

Tutore
Prof. Andrea Novelletto

Coordinatore
Prof. Fulvio Cruciani

TABLE OF CONTENTS

GLOSSARY

ALT = Alternative

bp = base pair

BWA = Burrows-Wheeler Aligner

CER = C-sites Enriched Region

CNV = Copy Number Variations

CO = Crossing Over

DBS = Double Strand Breaks

DHPLC = Denaturing High Performance Liquid Chromatography

DP = Depth

EMA = Exponential Moving Average

FilDP4 = Filter based on DP4

kya = kilo years ago

HJ = Holliday Junction

IR = Inverted Repeat

LINE = Long Interspersed Nuclear Element

Mb = Mega base pairs

MQ = Mapping Quality

MSY = Male-Specific region of the Y chromosome

mtDNA = Mitochondrial DNA

NAHR = Non-Allelic Homologous Recombination

NAGC = Non-Allelic Gene Conversion

NCO = Non-Crossing Over

NGS = Next Generation Sequencing

PAR = PseudoAutosomal Region

PSV = Paralogue Sequences Variant

REF = Reference

sd = standard deviation

SD = Segmental Duplication

SNP = Single Nucleotide Polymorphism

WGA = Whole Genome Amplification

SUMMARY

The Male-Specific region of Y chromosome (MSY) has long been considered a recombinationally inert genomic element, a view that has been profoundly dismissed by the discovery that the sequence landscape of this region can be actually modulated by inter- and intra- chromosomal recombination. The MSY includes eight large near-identical inverted repeats, palindromes P1-P8, considered to be in a ‘pseudo-diploid’ state. Although these structures originated in a non-recombining context, they evolved a strong self-recombinational activity in the form of non-allelic gene conversion, by which their arm-to-arm similarity exceeds 99.9%. Palindromic sequences contain many genes essential for sperm production and evolved independently on the constitutively haploid sex chromosomes of several taxa belonging to different kingdoms. Thus, it has been hypothesised that gene conversion is a mechanism necessary to counteract new possibly deleterious variants by retaining the ancestral state of gene sequences, in order to preserve their functionality over time. This hypothetical bias towards the retention of the ancestral state has also been proposed to explain the lower human-chimpanzee sequence divergence in the palindrome arms compared to palindrome spacers.

Despite the relevance of this mechanism in maintaining genome integrity, it is not clear if the bias towards the ancestral state really exists and little is known about the dynamics of gene conversion in the ampliconic MSY. Moreover, it is still unknown if differences in the conversion dynamics among different palindromes exist. Indeed, P8 palindrome showed evidence of X-to-Y gene conversion, but the interaction between Y-Y and X-Y recombination has never been clarified.

To shed light on these issues, we performed a high-depth (>50×) targeted next-generation sequencing (NGS) of palindromes P6, P7 and P8 in 157 samples, which cover the most divergent

evolutionary lineages of the human Y chromosome. We reconstructed a stable Y phylogeny of samples to explore gene conversion in an evolutionary context and found a large number of previously undescribed PSVs increasing the possibility to identify gene conversion events occurring during the recent human history. By mapping these events across our phylogeny and comparing the sequences of the three palindromes conserved between human and chimpanzee, we were able to infer the minimum number of conversions, palindrome-specific mutation and gene conversion rates and the direction of the gene conversion mechanism (ancestral/derived, GC-bias).

We found no evidence for a bias towards the retention of the ancestral state and that the evolution of different palindromes is governed by independent and complex dynamics.

We also observed higher mutation rates in the spacers compared to palindrome arms. This difference in mutation rate may represent the true cause of the previously observed higher human-chimpanzee spacer divergence with respect to the arms, without the need to invoke a Y-Y recombination bias towards the ancestral state.

INTRODUCTION

Segmental duplications

Features and evolution of segmental duplications

The human genome shows a complex pattern of nearly-identical interspersed segmental duplications (SDs) (International Human Genome Sequencing Consortium et al. 2001), that exist in multiple locations of the genome as a result of duplication events. SDs typically range in size from 1-200 kb and share high degree (90-100%) of sequence identity. The human genome has been found to be particularly enriched in these duplicates (about 5%) (Bailey et al. 2002) mapping to 2 or more points in the genome (Samonte and Eichler 2002). SDs have been associated with rapid gene innovation and chromosomal rearrangements in genomes of man and great apes and they appear to be crucial for evolution and disease (Bailey et al. 2002). In particular, mutation by duplication has two different consequences on an evolving genome: at first, through conservative transposition, it can lead to the birth of novel genes and regulatory elements; secondly, due to its high sequence identity, the duplication provides a substrate for subsequent rearrangements through the process of non-allelic homologous recombination (NAHR) (Carvalho and Lupski 2016). This dynamic mutational pattern may further increase the subsequent rounds of duplications as a result of abundant tracts of identical sequences (Dennis and Eichler 2016).

SDs may be distinguished in intra- and inter-chromosomal duplications (Eichler 2001). Their ancestral reconstruction has revealed a highly non-random organization with respect to their chromosomal distribution: intra-chromosomal SDs exhibit a higher degree of similarity and resulted to be more abundant compared to inter-chromosomal ones (Zhang et al. 2004; Sainz et al. 2006). In addition, it has been noted that across the genome, the Y chromosome exhibits the highest content of these highly similar duplicates (Sainz et al. 2006) (Figure 1).

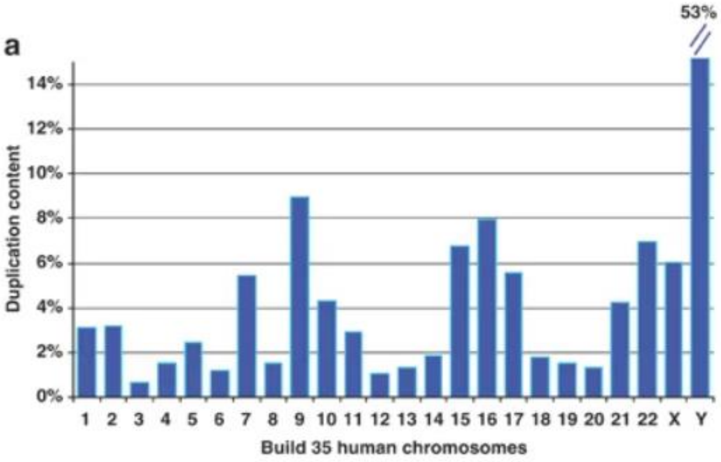

Figure 1: Relative percentage content of both intra- and inter-chromosomal segmental duplications among 24 human chromosomes. Each chromosome is represented by a column chart. The highest content is recorded within the Y chromosome (53%) (from Sainz et al. 2006).

It has also been found that among the intra-chromosomal SDs, the inverted repeats (IRs) are more abundant than direct ones (Sainz et al. 2006). Indeed, from the human genome wide-analysis of IRs emerged that out of the 47 IRs longer than 8 kb and showing >99% similarity, about 65% were located on the sexual chromosomes (Warburton et al. 2004), despite they only represent 6.5% of the whole genome. The abundance of these near-identical duplicates in sex-chromosomes represents a common feature not only of several primates, but also of more distantly related taxa (Ross et al. 2005; Ezawa et al. 2006; Davis et al. 2010; Geraldes et al. 2010; Soh et al. 2014; Skinner et al. 2016; Shi et al. 2019; Swanepoel et al. 2020; Zhou et al. 2020) and their independent appearance in different species suggests a potential functional and evolutionary role (Trombetta and Cruciani 2017).

SDs evolve thanks to several mutational processes, but due to their high sequence identity, they may undergo to homology-driven processes, such as NAHR (Bailey and Eichler 2006). Depending on the orientation of duplicates on the chromosomes, NAHR between two SDs may result in different structural re-arrangements, such as duplications and deletions involving the recombining fragment, or may give rise to inversion/translocation events (Figure 2A); but, if the molecular intermediates of NAHR are not resolved by crossing-over (Figure 2A), an alternative homology-driven scenario in represented by gene conversion, a recombining mechanism that mediates the non-reciprocal transfer of a short sequence from one SD copy to another (Figure 2B). As a consequence, gene conversion can increase the diversity at allelic copies while causing the homogenization of the paralogous SD copies, in a process known as concerted evolution.

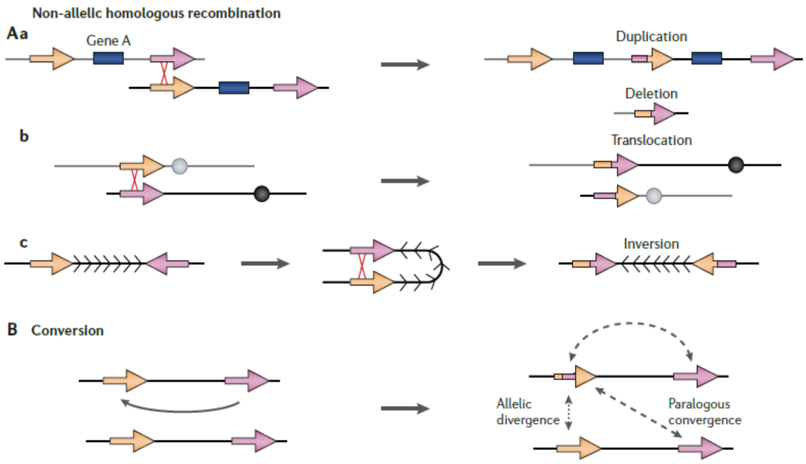

Figure 2: Homology-driven forces that affect the evolution of segmental duplications. **A)** Non-allelic homologous recombination (NAHR) may lead to different structural re-arrangements depending on the position and orientation of duplicates on the chromosomes: the crossing-over between two non-allelic direct SDs may originate duplication and deletion events (a); the NAHR between inter-chromosomal SDs may originate a reciprocal translocation (b), whereas the non-allelic crossing-over between two inverted intra-chromosomal repeats may give rise to an inversion (c). **B)** Gene conversion between two SDs as the result of the genetic information exchange between copies (from Bailey and Eichler 2006).

Genomic instability caused by NAHR in SDs has been found responsible of several disorders (Pentao et al. 1992; Stankiewicz and Lupski 2002; Conrad and Antonarakis 2007; Wang et al. 2018; Woodward et al. 2019). In addition, the mechanism of recombination that drives the genetic re-arrangements of SDs may have a crucial role in the molecular evolution of the human genome (Stankiewicz et al. 2004; Bailey and Eichler 2006; Estivill and Armengol 2007; Jiang et al. 2007; Yuan et al. 2015), as well as in the lineage-specific divergence between evolving species (Newman et al. 2005; Ventura et al. 2011).

The genetic diversity of SDs is linked to two different aspects: 1) the variation in the copy number of paralogous sequences (Copy Number Variants, CNVs) due to NAHR (Sharp et al. 2005) and 2) the sequence variation between two non-allelic SDs, represented by a Paralogue Sequence Variant (PSV), i.e. single base difference existing between paralogs due to point mutations occurring after the duplication.

In this study we will focus on the variability caused by PSVs, that represents the unique source of diversity between two near-identical paralogs.

Gene conversion

The Gene Conversion (GC) is a homologous recombination process by which one DNA sequence replaces a highly similar sequence such that the two DNA sequences become identical after the conversion event. Gene conversion can be either allelic or ectopic (non-allelic) (Duret and Galtier 2009). The first occurs between two alleles of a locus during meiotic recombination, the second occurs between two paralogous copies mapping at diverse points of the genome, either on the same or on different chromosomes (Galtier et al. 2001; Innan and Kondrashov 2010; Trombetta and Cruciani 2017).

In eukaryotes, gene conversion represents the main form of homologous recombination (considering both the allelic and non-allelic one) which starts from a DNA double-strand breaks (DSBs), caused by a DNA damage. Unlike the crossing-over (CO), its mechanism involves the non-reciprocal transfer of genetic information from an intact “donor” sequence to a broken “acceptor” one (Chen et al. 2007).

The mechanism of gene conversion

The mechanism of gene conversion, outlined in Figure 3, starts from a DSB of the DNA, which can be induced by a topoisomerase-like enzyme during meiosis or by radiation, stalled replication forks and specific endonucleases, during mitosis.

After the formation of the DSB, the 5’ extremities of the broken strands are degraded by 5′→3′ exonucleases, resulting in the formation of two 3′ single stranded DNA tails (Chen et al. 2007). After scanning the genome, a 3’ ssDNA segment invades a homologous sequence, forming an intermediate displacement (D-loop) which can be repaired through different pathways, resulting either in crossover or non-crossovers (NCO). The D-loop is extended by DNA syintetis, then it invades the other 3′ ssDNA tail, and the ligation of nicks results in an intermediate characterized by two Holliday-junctions (HJs) (Figure 3b).

The resolution of HJs by a HJ resolvase is possible through both crossover and gene conversion (NCO) (Figure 3d), but during this process the gene conversion resulted to be the most probable repairing process of mismatches (Szostak et al. 1983), where the correction of the broken strand occurs using the intact strand as a template (Haber et al. 2004).

This model explains the resolution of DSBs during meiotic recombinatin, whereas the highest frequency of gene conversion compared to CO (< 8%) in solving the induced DSBs of the DNA (Ira et al. 2006), is explained by other two models: 1) the Synthesis-Dependent Strand-Annealing (SDSA) (Figure 3c) and 2) the double-HJ dissolution (Figure 3e).

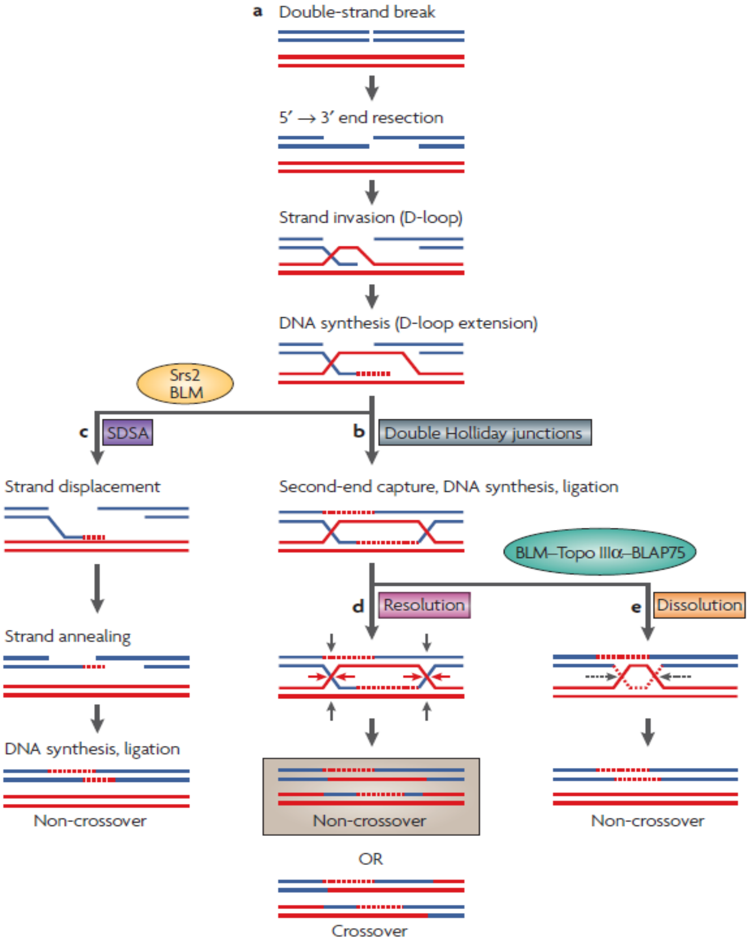

Figure 3: Mechanism of resolution of DSBs of the DNA by gene conversion. a) double strand breaks of the DNA, b) Double Holliday junctions formation, c) SDSA model, d) Resolution of Holliday junctions by gene conversion or crossover, e) Double-HJ dissolution model (from Chen et al. 2007).

In the SDSA model, after the D-loop extension, the newly synthesized strand is displaced from the template and anneals to the other 3′ ssDNA tail; this is followed by DNA synthesis and ligation of nicks. From this model, only non-crossover products are generally yelded (Szostak et al. 1983). According to the double-HJ dissolution model, non-crossover products are generated from the convergent migration of the two HJs towards each other, leading to the collapse of the double HJs (Chen et al. 2007).

Features of non-allelic gene conversion

The non-allelic gene conversion (NAGC) is a kind of NAHR where the copying of the genetic information from the donor sequence to the acceptor occurs between paralogs which are at distinct genomic loci (Harpak et al. 2017) (Figure 4), on the same or different chromosomes. It is possible when paralogous sequences are accidently aligned during recombination because of their high similarity (Chen et al. 2007). It has been proposed that an efficient gene conversion requires at least 88% homology between interacting blocks (De Marco et al. 2000), even if it efficiently occurs when sequences display >95% similarity (Chen et al. 2007). It has been proposed that the frequency of the recombination is inversely proportional to the distance between the interacting loci (Schildkraut et al. 2005).

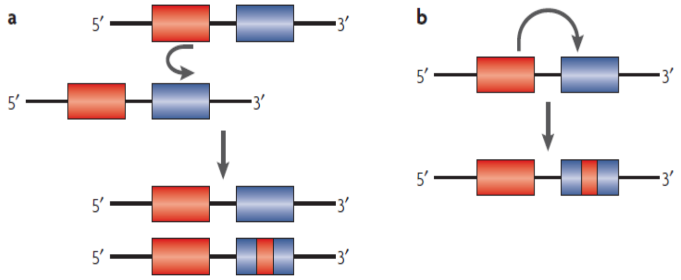

Figure 4: Non-allelic gene conversion. **a)** NAGC in trans, shown as an event occurring between paralogous sequences located on sister chromatids or on homologous chromosomes. **B)** NAGC events in cis, occurring between non-allelic loci that reside on the same chromosome (adapted from Chen et al. 2007).

Although gene conversion tracts are usually long in yeast (Mancera et al. 2008), in mammals they range from 200 bp to few kilobases. Some examples are given by the estimates for the human globin genes, which exhibit conversion tracts of 113-2,266 bp (Papadakis and Patrinos 1999) or a more limited range of 1-1,365 bp observed for human endogenous retroviral (HERV) sequences of the Y chromosome (Bosch et al. 2004). Interestingly, a tract of 9,023 bp has been found within palindrome P6 of the human Y chromosome by Hallast and colleagues (2013). However, the Y-Y gene conversion tract rarely exceeds 1 kb, nevertheless it results to be usually longer than the converted tract in the X-to-Y NAGC (Bosch et al. 2004; Rosser et al. 2009; Cruciani et al. 2010b; Trombetta et al. 2010, 2014).

The possibility to estimate the maximum length of a NAGC event is given by the observation of a set of co-converted adjacent PSVs delimited by the two nearest non-converted PSVs (Hallast et al. 2013).

There is evidence that the NAGC is a biased process: the rate of conversion from one paralog to another may be higher than the rate of the reciprocal transfer (Bosch et al. 2004; Chen et al. 2007). Moreover, when a PSV exists, a preferential direction in the retention of specific variants has been observed, favouring some allels over others (Marais 2003). After a DSB of the DNA, the repair of A:G and C:T mismatches preferentially retains G and C bases over the other two alleles (Chen et al. 2007). This results in a biased Gene Conversion (bGC) towards G or C alleles and it has been observed for both allelic and non-allelic processes (Galtier 2003; Kudla et al. 2004). The main effect of the bGC is to increase the GC content of the sequences where it occurs (Hallast et al. 2013).

Evolutionary and functional consequences of NAGC

The non-allelic gene conversion does not affect identical sites between paralogous sequences, and it is only detectable when a PSV exists between them (Hallast et al. 2013).

The key role of NAGC is to maintain the high level of sequence identity between SDs by eliminating differences between the two interacting paralogs. For this reason, it has been found to be implicated in many cases of concerted evolution of human gene families (Verrelli and Tishkoff 2004; Hallast 2005; Woelk et al. 2007), with the consequence that NAGC may alter the evolutionary relations between SDs, making the paralogous gene sequences more closely related to each other than they are to their orthologous counterparts in closely related species (Rozen et al. 2003) (Figure 5).

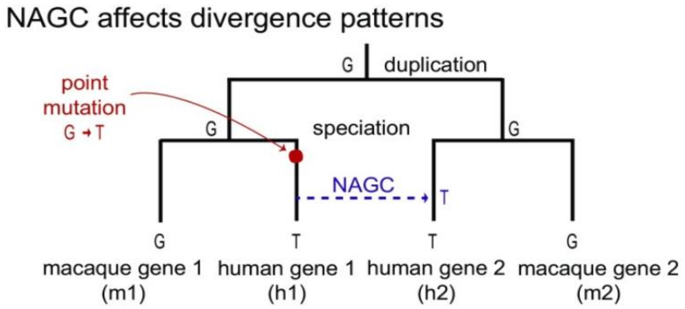

Figure 5: The NAGC can drive rare divergence patterns, like the sharing of alleles between paralogs but not orthologous sequences (adapted from Harpak et al. 2017).

One of the most relevant example of the concerted evolution is represented by the multi-copy gene families that lie within the eight large palindromes of the euchromatic portion of the male specific region of the human Y chromosome (MSY), where the NAGC works in order to repair the DNA damage of essential genes to preserve their functionality over time (Rozen et al. 2003).

If on one hand the NAGC acts by enhancing the homology between paralogs, on the other hand it may affect the genetic variability of a genome by increasing the allelic diversity within the population (Figure 6), since the donor paralog acts as a reserve of variability (Chen et al. 2007). When a gene conversion event occurs between two paralogs holding a variant (Figure 6A), the PSV will disappear, but a new SNP will be introduced in the population of chromosomes (Trombetta and Cruciani 2017) (Figure 6B). If an abundant gene conversion occurs, its result is to eradicate differences between paralogs (PSVs) and among chromosomes (SNPs) (Trombetta and Cruciani 2017) (Figure 6C).

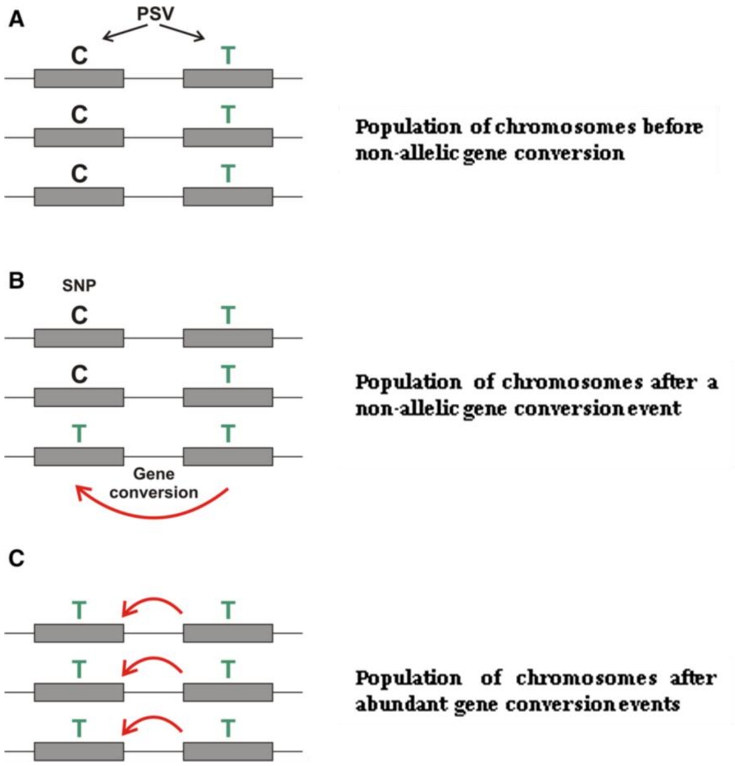
A
PSV
C
T
C
T
C
T
Population of chromosomes before non-allelic gene conversion
B
SNP
C
T
C
T
T
T
Gene conversion
Population of chromosomes after a non-allelic gene conversion event
C
T
T
T
T
T
T
Population of chromosomes after abundant gene conversion events

Figure 6: Effect of inter-paralog gene conversion within a population of chromosomes. **A)** Paralogue sequence variant between SDs in a population of chromosomes. **B)** A single inter-paralogue gene conversion event eliminates differences between paralogs of a chromosome and creates a new SNP among chromosomes. **C)** Several inter-paralogue gene conversions may eliminate differences between SDs (PSVs) and chromosomes (SNPs) (adapted from Trombetta and Cruciani 2017).

The majority of disease-associated mutations are constituted by single-base substitutions and short deletions/insertions, resulting from DNA replication errors. However, pathological mutations can also be introduced by non-reciprocal recombination events between disease-associated genes and their paralogous sequences (Casola et al. 2012). The non-allelic gene conversion events between SDs were also indicated as the molecular cause of a large number of human inherited diseases (Chen et al. 2007); in fact, if the donor sequence is not functional, such as a pseudogene, it can accumulate mutations escaping the negative selection. In this way, these new variants can convert the active functional paralogs by inhibiting the activity of the receiving genes (Surdhar et al. 2001; Vanita et al. 2001; Nakashima et al. 2004; Bischof et al. 2006; Friães et al. 2006; Casola et al. 2012). In light of these events, it is crucial to characterise the dynamics of inter-paralogue gene conversion within human genome.

It may be really difficult to understand the dynamics of NAGC by performing the comparative sequence analysis of paralogs among different genomes, since the diploidy of the human genome does not make possible to distinguish between PSVs and allelic variants. In addition, the homologous recombination between allelic loci may further complicate the interpretation of the ancestral status of the variants. In order to overcome these limitations, in this study we will examine the non-allelic gene conversion within the haploid MSY, which is particularly enriched in SDs, called palindromes. Moreover, thanks to its haploid features and the lack of meiotic recombination, the reconstruction of haplotypes based on rare SNPs of non-duplicated regions of the MSY is helpful to reconstruct a non-ambiguous phylogeny and to infer the ancestral status of variants by mapping them in this evolutionary context (Karafet et al. 2008).

The human Y chromosome

The Y chromosome evolution

The mammalian sex chromosomes are thought to have arisen from an ordinary pair of autosomes about 300 million years ago, when an autosome acquired a mutation that gave rise to a dominant allele in the masculinity region, the testis-determining factor (TDF), conferring an enormous reproductive advantage to the individuals who possessed it (Graves 2006). The acquisition of the SRY gene (sex determining region of Y) for sex determination represented the starting point for the differentiation of the proto-Y from its homologous counterpart. Then, the natural selection acted in order to avoid the loss of this sex-specific benefit by disadvantaging mutations within SRY and favouring the accumulation of other genes with male-specific functions near the locus (Figure 7). In this context, the subsequent suppression of recombination limited the possibility of repairing DNA mutations, so that the proto-Y underwent a series of deletions causing its progressive shortening, in addition to the accumulation of a series of nucleotide substitutions, inversion and insertions, which led to an increasingly evident differentiation between sex chromosomes (Graves 2006).

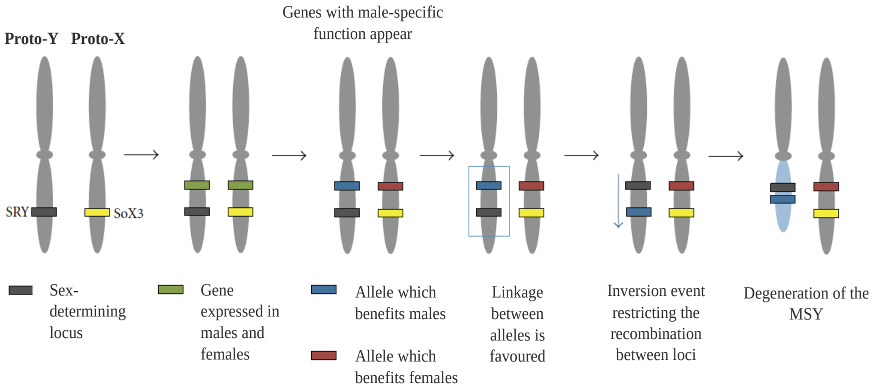
Genes with male-specific function appear
Proto-Y
Proto-X
SRY
SoX3
Sex-determining locus
Gene expressed in males and females
Allele which benefits males
Allele which benefits females
Linkage between alleles is favoured
Inversion event restricting the recombination between loci
Degeneration of the MSY

Figure 7: The model of sex chromosomes evolution. The SRY gene acquisition represent the starting point for the Y chromosome morphological differentiation from the X chromosome (adapted from Betrán et al. 2012).

The suppression of recombination occurred in 5 different evolutionary steps, each of which is the result of a large inversion (Lahn and Page 1999; Ross et al. 2005). These events resulted in the formation of 5 evolutionary layers in the X-degenerate region, which exhibit different homology with the paralogous counterpart on the X chromosome (60-96%), reflecting the 5 moments of the interruption of recombination, occurred between 240 and 30 million years ago. Then subsequent structural rearrangements led to the loss of physical continuity between elements of a layer (Ross et al. 2005).

Another evolutionary mechanism involving the Y chromosome is the accumulation of genes from other chromosomes and the duplication of these or other genes already present on the Y. About the 25% of the MSY is composed of ampliconic sequences, which are believed to be originated by duplication of sequences of autosomal origin (DAZ and CDY) or belonging to the X-degenerate region (RBMY and VCY genes) (Skaletsky et al. 2003). Other genes possibly originated de novo on the Y (PRY and BPY2) since no X-linked or autosomal homologues have been identified (Betrán et al. 2012).

Among the evolutionary mechanisms that led to the current structure of Y chromosome, there are gene conversion events that affect the ampliconic region. Indeed, after the gene duplication, the NAGC acted in order to maintain the high similarity between copies. A possible explanation for this event consists in the fact that most of the genes of the MSY ampliconic region have a tissue-specific expression in the testes (germ line), and this unique recombination form is necessary to maintain the structural integrity of fertility genes and their functions during evolution, in absence of crossing-over with a homologous chromosome (Skaletsky et al. 2003).

Structure of the human Y chromosome

The human Y chromosome is ~59 Mb long and is one of the smallest chromosomes of the human genome, accounting for less than 2% of the haploid genome (Morton 1991; International Human Genome Sequencing Consortium 2001; Skaletsky et al. 2003). Its extremities are composed by two pseudo-autosomal regions, PAR1 and PAR2 (Pseudo-Autosomal Region 1 and 2), which in total represent the 5% of the chromosome length, being, respectively 2.7 Mb (Cotter et al. 2016) and 330 kb (Freije et al. 1992; Ross et al. 2005). These two terminal blocks exhibit a high recombination activity with the allelic portions of the X chromosome, thus genes located in these regions are present in two copies in both male and female and are inherited as autosomal genes (Page et al. 1987; Ross et al. 2005).

Genomic studies revealed that the Y chromosome contains a region comprising 95% of its length, where the X-Y crossing-over does not occur. This region was initially referred to as the Non-Recombining region of Y (NRY) (Blanco et al. 2000), but the discovery of an abundant form of recombination occurring among the IRs of the Y chromosome led the scientists to rename it the Male-Specific region of Y, or MSY (Rozen et al. 2003; Skaletsky et al. 2003; Hallast and Jobling 2017; Jobling and Tyler-Smith 2017; Trombetta and Cruciani 2017). The MSY shows a male uniparental inheritance. It is made up of both euchromatic and heterochromatic fractions (Figure 8). The heterochromatic region consists of three blocks of sequences that are enriched in tandem short repeated sequences and shows no evidence of transcription activity (Skaletsky et al. 2003).

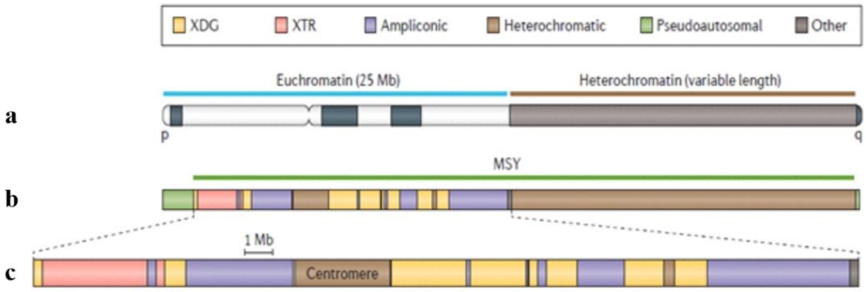

Figure 8: The structure of the human Y chromosome. **a)** Schematic representation of the Y chromosome, including PARs and heterochromatic region. **b)** MSY classes of sequences. **c)** Enlargement of the MSY euchromatin. The three classes of sequences are shown, as well as the heterochromatic fraction (adapted from Jobling and Tyler-Smith 2017).

The euchromatic region is roughly 22.5 Mb long, including 8 Mb on the short arm (Yp) and 14.5 Mb on the long arm (Yq). The euchromatin is characterized by the presence of 156 transcriptional units, 78 of which are protein-coding, and 60 of them are members of nine gene families, each characterized by more than 98% nucleotide identity among family members. The remaining 18 protein-coding genes are present in single copy in the MSY (Skaletsky et al. 2003).

The euchromatic MSY may be considered as a mosaic of three discrete classes of sequences, named X-transposed, X-degenerate and ampliconic regions.

The X-transposed region (XTR) arose by a single transposition of material from the X to the Y chromosome since the divergence of human-chimpanzee lineages (~4.7 MYA) (Page et al. 1984; Skaletsky et al. 2003; Ross et al. 2005), indeed it shares 99% of identity with the X chromosomal band Xq21. Then, an inversion within the MSY short arm led to the separation of XTR into two non-contiguous blocks (Figure 8). It is characterized by a 3.4 Mb combined length and hosts two coding-genes, together with a lot of interspersed repeated elements, such as LINE1 (long interspersed nuclear elements 1). Differently from PARs, the X-transposed region does not participate in crossing-over during male meiosis.

The X-degenerate region (XDG) mainly corresponds to the evolutionary relicts of the homologous regions of the ancestral autosomes, from which the sex chromosomes evolved (Lahn and Page 1999; Lahn et al. 2001; Skaletsky et al. 2003; Ross et al. 2005; Graves 2006; Jobling and Tyler-Smith 2017). It exhibits from 60% to 96% sequence identity with the corresponding paralogous sequences of the X chromosome, and is characterized by the presence of single copy genes or pseudogenes which share homology with 27 genes on the X. It includes almost all MSY genes that are ubiquitously expressed in the organism, and the SRY gene, responsible for the sex-determination, exclusively expressed in the testes (Skaletsky et al. 2003).

Finally, the ampliconic region has a 10.2 Mb combined length and represents about 1/3 of the euchromatic MSY. It mainly consists of intrachromosomal SDs organized in 8 inverted repeated sequences, called palindromes, which exhibit a marked similarity (>99.9%) (Skaletsky et al. 2003).

Human MSY palindromes

The MSY palindromic sequences are the most peculiar structures of the ampliconic region, they are located on the long arm of the Y chromosome and are designated as P1-P8 (Figure 9) basing on their order with respect to the Yq telomere. Palindromes are made up of two inverted repeats, called arms, separated by a non-duplicated spacer, and span in total for 5.7 Mb, amounting to 25% of the MSY euchromatin (Rozen et al. 2003; Skaletsky et al. 2003; Trombetta and Cruciani 2017). Due to their duplicated nature, palindromes exhibit 'pseudo-diploid' features.

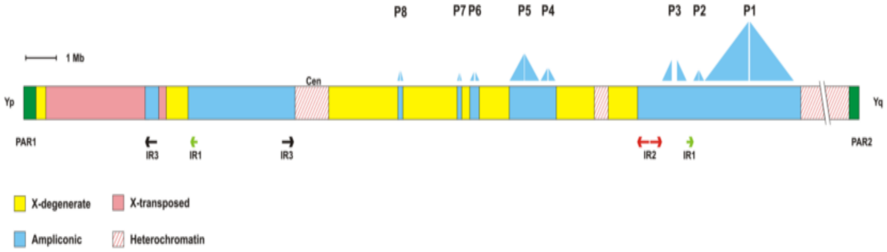

Figure 9: Localization of MSY palindromes along human Y chromosome. Triangles denote the inverted repeated arms of the eight palindromes (P1–P8), while gaps between triangles represent the non-duplicated spacers. At the bottom are shown the three inverted repeats (IR1–IR3) with a sequence identity above 99.6% (adapted from Trombetta and Cruciani 2017).

There is no correlation between arms and spacer length. To distinguish between palindrome arms in this study, the arm located closest to the centromere will be referred to as the ‘proximal’ arm, whereas the arm situated further from the centromere will be referred to as the ‘distal’ palindrome arm. The highly symmetrical arms exhibit an arm-to-arm nucleotide identity ranging from 99.94% to 99.997% (Table 1), only referring to the single nucleotide substitutions represented by PSVs. In addition to palindromes, within the MSY ampliconic region have been identified 5 long inverted repeats. The IRs are similar in structure to palindromes, but contain much larger spacers and exhibit lower sequence similarity between paralogs (95-99.95%) (Skaletsky et al. 2003; Trombetta and Cruciani 2017).

Palindromic sequences are enriched in active coding genes organized in gene families (Skaletsky et al. 2003; Hughes et al. 2010) (Table 1). Six out of the eight palindromes carry recognized protein-coding genes that are expressed specifically in testes, each of them having an identical gene copy on the opposite arm of the palindrome. Of the nine multi-copy gene families identified in the MSY, six are located exclusively in palindromes (VCY, XKRY, CDY, HSFY, PRY and DAZ), whereas the other two have members also on palindromes (BPY2 and RBMY). In addition, palindromic sequences contain at least seven families of apparently non-coding transcription units, expressed exclusively or predominantly in testes (Skaletsky et al. 2003).

Pal	**Arm length (kb)**	**Spacer length (kb)**	**% Identity**	**Genes**
P1	1,450	2.1	99.97	DAZ, CDY, BPY2
P2	122	2.1	99.97	DAZ
P3	283	170	99.94	RBMY, PRY
P4	190	40	99.98	HSFY
P5	496	3.5	99.98	CDY, XRKY
P6	110	46	99.97	None
P7	8.7	12.6	99.97	None
P8	36	3.4	99.997	VCY

Table 1: MSY palindrome features according to data reported in Skaletsky et al. (2003).

Advances in DNA sequencing technologies, together with the increasing affordability of large-scale sequencing projects, have dramatically expanded opportunities for interspecific comparative genomics (Hughes et al. 2005, 2010, 2012; Soh et al. 2014; Tomaszkiewicz et al. 2016; Cechova et al. 2020). In particular, palindromic sequences have also been found in Gorillas and Chimpanzees, so it was thought that they originated before the evolutionary separation between humans and great apes, but due to the conservation of human P4 and P5 palindromes in macaque ampliconic region, it has been hypotesized a deeper origin of these structures, tracing their origin back to over 25 MYA (Hughes et al. 2012; Cechova et al. 2020). All human palindromes, with the exception of P3 and P4, have orthologs in the chimpanzee MSY, which also contains other 12 species-specific palindromes (Hughes et al. 2010).

The high similarity between palindrome arms is though to be explained by repeated non-allelic gene conversion events, necessary to preserve the high sequence identity among genes of the ampliconic region (Rozen et al. 2003; Skaletsky et al. 2003; Hallast et al. 2013, Trombetta and Cruciani 2017; Skov et al. 2017; Hallast and Jobling 2017; Jobling and Tyler-Smith 2017), whose mutation can cause male infertility (Ali and Hasnain 2003; Dhanoa et al. 2016). In addition, the human-chimpanzee sequence divergence within palindrome arms has been found to be significantly lower than divergence within spacers (Rozen et al. 2003), suggesting that Y-Y gene conversion since speciation must have been directional, tending to revert mutations arising in the arms back to their non-mutated state. This finding, together with the possibility to preserve gene sequences from the evolutionary decay, led to the hypothesis that gene conversion is a mechanism biased towards the retention of the ancestral state of variants (Rozen et al. 2003; Hallast et al. 2013; Skov et al. 2017).

Human X and Y chromosomes are particularly enriched in inverted repeats (Warburton et al. 2004), that also are a common feature of sex-limited chromosomes of many distant taxa, suggesting that their evolution may be linked to essential biological functions.

Characteristics of P6, P7 and P8 palindromes

Within the human MSY, palindromes are not identical each other, some of them show peculiar and more complex structural features than others. In particular, P1 hosts two secondary palindromes (P1.1 and P1.2) and its central part is nearly identical to the adjacent P2 palindrome (Kuroda-Kawaguchi et al. 2001). Other portions of P1 share homology with P3, P4 and P5, as well as to other non-palindromic sequences (Skaletsky et al. 2003; (Bhowmick et al. 2007; Costa et al. 2008). Non-allelic recombination can also occur among paralogous regions of different palindromes, resulting in duplications, deletions and inversions. Thus, in order to avoid further complication of ‘pseudo-diploid’ regions analysis, in this study we will focus exclusively on singleton palindromes of the MSY: P6, P7 and P8, characterised by a single repeat unit for each arm. In addition, singleton palindromes are commonly found on mammalian sex chromosomes and allow more accurate identification of PSVs and gene conversion events.

P6, P7 and P8 are the three shortest palindromes of the human MSY (Table 1) and all of them are evolutionary conserved in chimpanzee. Palindromes are hypothesized to be evolved to allow ampliconic genes to withstand high mutation rates on the Y via gene conversion, in the absence of interchromosomal meiotic crossing-over. However, P6 and P7 palindromes (266 and 30 kb, respectively) do not harbour any known protein-coding genes (Skaletsky et al. 2003; Chechova et al. 2020), but despite this, they have been found to be present in multi-copy state in the human-gorilla common ancestor (Cechova et al. 2020). P8 palindrome is characterized by a slightly more complex structure: its arms are

38.0 and 37.4 kb long (proximal and distal arm, respectively) (Shi et al. 2019), also including two additional ~2.8 kb and ~2.2 kb low-similar paralogs flanking the arms. These two paralogous fragments harbour a ~0.6 kb difference in the reference sequence. Each arm hosts a VCY gene copy, and it is one of the palindromes that has a similar structure size to chimpanzee (Hughes et al. 2010). Nevertheless, it was shown to be variable in copy number among samples analysed in the 1000 Genomes Project phase 3 (Poznik et al. 2016; Teitz et al. 2018).

The VCY genes are members of a gene family also comprising four X-chromosomal members, designated as VCX, VCX2, VCX3A and VCX3B. These gene copies are placed within four paralogous sequences of the X chromosome (gametologs), spanning from ~10 to ~16 kb, and similarly to the PSVs, the single base differences between X and Y gametologs are refferred to as Gametologous Sequence Variants (GSVs). The VCX/Y genes appear to be expressed exclusively in male germ cells and encode small, positively charged proteins (Lahn and Page 2000). Members of the VCX/Y family share a high degree of sequence identity (>98%) (Ross et al. 2005), with the exception that a 30 nucleotide unit is tandemly repeated in X-linked members but is present only once in Y-linked copies (Lahn and Page 2000). The high similarity between X and Y copies is probably due to the effect of X-Y inter-chromosomal gene conversion (Trombetta et al. 2010).

VCY proteins have been largely detected in germ cell nuclei (Zou et al. 2003), and VCX3A gene has been previously proposed as the candidate gene for X-linked mental retardation and icthyosis (Hosomi et al. 2007; Ben Khelifa et al. 2013). However, the biological function of the testis-specific family members in the whole organism is still unknown. Desptite this, VCX/VCY genes exhibits evidence of a strong X-to-Y gene conversion (Cruciani et al. 2010b; Trombetta et al. 2010), making P8 palindrome being involved in both intra and inter-chromosomal NAGC.

Genetic variability of the Y chromosome

Biallelic polymorphisms

A polymorphism is defined as a genetic variation with the minor allele present at a frequency of at least 1%. We refer to biallelic polymorphisms when a variant occurs in two possible allelic states: "ancestral" and "derived".

With the introduction of the Denaturing High Performance Liquid Chromatography (DHPLC) and the improvement of the Sanger sequencing technique, it has been possible to discover several hundreds of new biallelic polymorphisms within the MSY, including SNPs (Underhill et al. 1997, 2000, 2001; Shen et al. 2000, 2004; Cruciani et al. 2002, 2004, 2006, 2007, 2008, 2010a; Hammer and Zegura 2002; Y Chromosome Consortium 2002; Kayser et al. 2006; Mohyuddin et al. 2006; Underhill and Kivisild 2007; Karafet et al. 2008; Chiaroni et al. 2009; Cruciani et al. 2011; Trombetta et al. 2011; Scozzari et al. 2012,; Mendez et al. 2013), whereas in recent years, the advent of new generation sequencing techniques made possible to proceed faster in the high-resolution analysis of the genetic variability of the Y chromosome, resulting in the identification of thousands of SNPs (Xue et al. 2009; 1000 Genomes Project Consortium et al. 2010; Francalacci et al. 2013; Poznik et al. 2013; Wei et al. 2013; Scozzari et al. 2014; Hallast et al. 2015; Karmin et al. 2015; Trombetta et al. 2015a; Barbieri et al. 2016; Poznik et al. 2016; D’Atanasio et al. 2018; Finocchio et al. 2018; Grugni et al. 2019; Rivera Franco et al. 2020).

From these studies emerged that the genetic variability of the MSY is lower than that observed in the X chromosome, autosomes and mitochondrial DNA (mtDNA) (Sachidanandam et al. 2001). In addition to the hypothesis of a very recent common ancestor for the Y chromosomes, the lower diversity of the MSY is explained by 1) the small effective population size (Hammer 1995; Underhill et al.

1996), which is four times smaller than that of autosomes and three times smaller than that of the X chromosomes, making the Y chromosomes more susceptible to the effects of genetic drift; and 2) the MSY behaves as a unique linkage group, because of the absence of meiotic recombination, leading to the fixation of a combination of alleles which are under positive selection (Rice 1987; Whitfield et al. 1995).

Thanks to their low mutation rate ($<10^{-9}$ events/position/year) (Myres et al. 2011; Mendez et al. 2013; Fu et al. 2014; Scozzari et al. 2014; Helgason et al. 2015; Trombetta et al. 2015b; D'Atanasio et al. 2018; Finocchio et al. 2018) the biallelic polymorphism may be considered stable in evolutionary time. Thus, if we observe two or more Y chromosomes showing the derived status at the same biallelic site, it is likely that they derive from the same common ancestor. According to this, chromosomes showing the derived status at polymorphic sites can be grouped into monophyletic entities called haplogroups, that may be arranged in a unique and stable phylogeny (Karafet et al. 2008).

Multiallelic polymorphisms

The Y chromosome hosts several classes of multiallelic polymorphisms, represented by minisatellites, microsatellites, telomeric repeats (etc.).

The most important class used in research field is represented by microsatellites, which consist in short tandem repetitions of 1-6 bp, with a high level of polymorphism represented by different number of repetitions of the pattern (Jobling et al. 2013). Microsatellites generally mutate by acquisition or loss of a single repeated unit (Weber and Wong 1993; Di Rienzo et al. 1994; Kayser et al. 2000; Kayser et al. 2004; Gusmão et al. 2005), and the accepted model for the generation of new alleles is the "slipped strand mispairing" (Levinson and Gutman 1987), consisting in an incorrect pairing of the repetitive units of the microsatellite caused by the slipping of a DNA strand during the replication event.

The average mutation rate of microsatellites is approximately 2 × 10^{-3} mutations/bp/generation, several orders of magnitude higher than that observed for SNPs. For this reason, the equality by state of different chromosomes could be due to several independent mutational events instead of identity by descent. For this reason, these markers are not suitable to reconstruct phylogenetic trees.

Copy number variations

The Copy Number Variations (CNVs) includes both biallelic and multiallelic polymorphisms; this class refers to variations in the copy number of DNA fragments longer than 1 kb (Feuk et al. 2006).

The variation in the number of copies of a sequence may be the result of different mechanisms involving homologous sequences, such as SDs, or may occur in proximity of regions where homology is limited to 2-15 bp (micro-homology). Many of these mechanisms have the physiological role of repairing single or double-strand breaks of the DNA, but may also result in the alteration of the chromosomal structure and copy number of the affected sequence. The NAHR is the main mechanism by which the duplication or deletion of a locus occurs (Massaia and Xue 2017), as the result of the pairing in meiosis of paralogous sequences located on the same or on different chromosomes (unequal crossing-over).

Due to its haploidy, the Y chromosome accumulated a higher percentage of SDs compared to the rest of the genome; this led to the generation, especially through NAHR, of a high number of CNVs. These variants can be placed within a phylogenetic tree of the Y chromosome, and through the definition of their ancestral or derived state, it is possible to estimate the time needed for the generation of CNVs (Jobling 2008).

Biallelic markers and Y chromosome phylogeny

If we exclude the Y-Y gene conversion events between the intrachromosomal SDs of the ampliconic region (Rozen et al. 2003; Hallast et al. 2013; Balaresque et al. 2014; Trombetta et al. 2016; Skov et al. 2017; Trombetta and Cruciani 2017) and between X-Y SDs recently described (Rosser et al. 2009; Trombetta et al. 2010, 2014), the MSY lacks any other recombination event; thus the only source of the Y chromosome genetic variability is the sequential accumulation of new mutations (Rozen et al. 2003; Skaletsky et al. 2003). Over time, this process has given origin to monophyletic entities, called haplogroups (Jobling and Tyler-Smith 2003, 2017).

Haplogroups are defined by the derived allele of biallelic markers (usually SNPs), which, due to their low mutation rate, may be considered unique events during human evolution. Thanks to the MSY haploidy and lack of recombination, the combination of single alleles of the MSY defines stable haplogroups, which may be organized in an unambiguous phylogenetic tree.

The Y chromosome phylogenetic tree topology can be reconstructed from the typing of mutations in different human populations, and its structure changes when new mutations are detected. In 2008, new biallelic polymorphisms were discovered and grouped into a large high-resolution tree composed by 20 main clades, indicated from 'A' to 'T' (Figure 10) (Karafet et al. 2008). Each haplogroup is represented by a branch of the tree and each branch is associated with a name, according to a standard nomenclature system (Y Chromosome Consortium 2002).

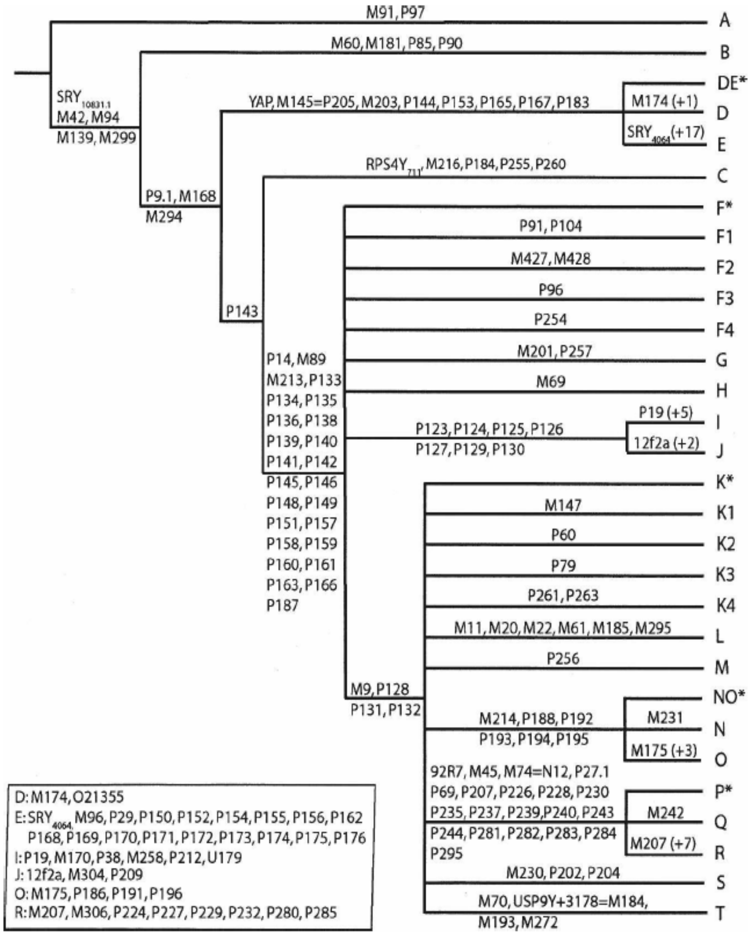

Figure 10: Main haplogroups of the Y chromosome phylogenetic tree. The Y phylogeny allows Y chromosomes to be assigned to a specific haplogroup based on the combination of allelic states for binary markers. Haplogroups are arranged in clades named from ‘A’ to ‘T’ and each clade may be subdivided into subclades. Mutation names are indicated along the branches (adapted from Karafet et al. 2008).

Since then, the next generation sequencing (NGS) led to the discovery of thousands of new SNPs, allowing the improvement of the resolution of the Y phylogenetic tree (Wei et al. 2013; Mendez et al. 2013; Scozzari et al. 2014; Trombetta et al. 2015a; Poznik et al. 2016; D’Atanasio et al. 2018).

Due to the molecular differentiation occurred during the colonization of different areas of the world by the human species, each haplogroup tends to be localized in one or few geographical areas. The geographical distribution of the main Y chromosome haplogroups is represented in Figure 11.

The most recent analysis of the world-wide geographical distribution of Y-chromosome haplogroups is that reported in Jobling and Tyler-Smith (2017), based on more than 60,000 SNPs identified by the low-coverage sequencing of 1,244 present-day chromosomes (Poznik et al. 2016).

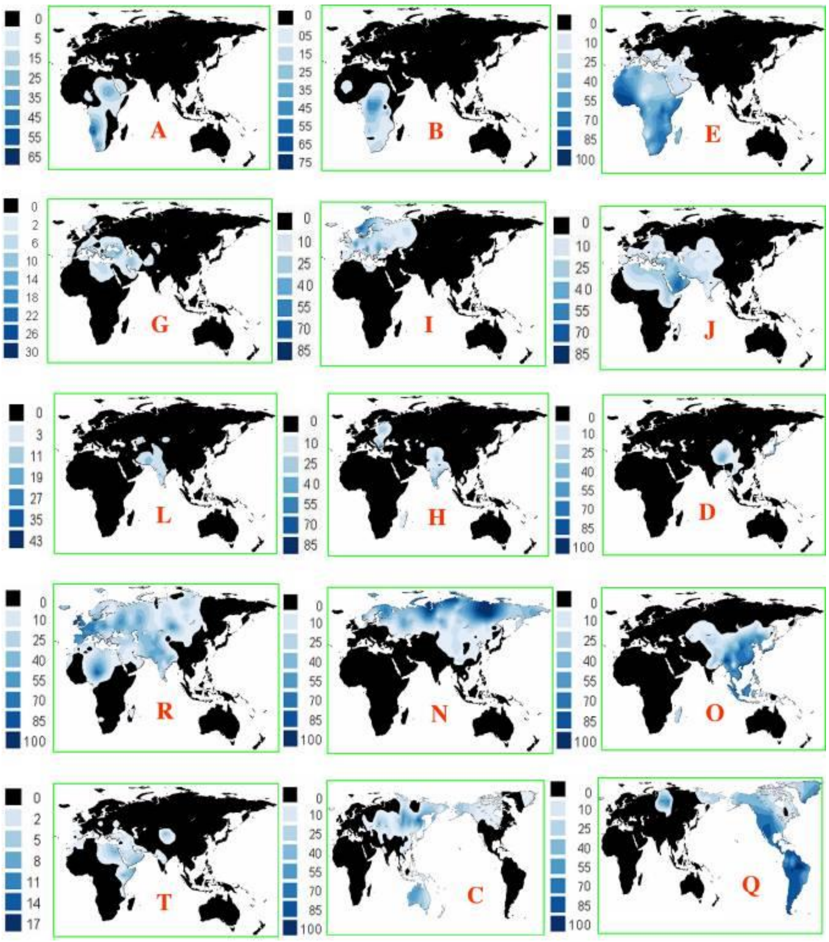

Figure 11: World-wide frequency distribution maps for the main MSY haplogroups (from Chiaroni et al. 2009).

Y chromosome phylogeny and gene conversion

The MSY corresponds to the genomic portion most covered by inter- and intra- chromosomal SDs, which provide plentiful substrates for frequent non-allelic gene conversion. This gives the possibility that the biallelic markers localized within duplicates, such as MSY palindromes, show unusual mutational properties which reflects the action of NAGC. Frequent gene conversion events between palindrome arms may alter their pattern of sequence identity together with the evolutionary relationships between chromosomes (Bosch et al. 2004). This is possible by introducing the derived state of a SNP on both paralogous arms, resulting in new mutated positions; whereas a gene conversion event directed to the ancestral state can erase the mutated state of a variant (Adams et al. 2006), masking, as a consequence, the mutational event. Abundant NAGC may also introduce the derived state of a SNP on different branches of the phylogenetic tree, that will be wrongly interpreted as recurrent mutations.

For these reasons, mutations occurring within SDs of MSY do not provide useful data for the comprehension of evolutionary phenomena and the reconstruction of a phylogenetic tree, making the biallelic markers of the unique regions much more suitable for this purpose. Despite this, palindromes cannot undergo independent evolutionary histories. When we observe a PSV determining a ‘pseudo-heterozygous’ state, such as G/A, then the observation in other chromosomes of the two other possible ‘pseudo-homozygous’ genotypes, G/G and A/A, indicates that gene conversion occurred during the evolution of the examined sequences, assuming that recurrent mutations can be ignored (Rozen et al. 2003; Hallast et al. 2013). However this does not clarify how many independent gene conversion events delineated the three genotypes, but, thanks to the availability of a stable phylogeny defined by binary markers falling outside palindromic regions, we are able to investigate the evolutionary relationships established among chromosomes of different Y lineages and precisely determine how many gene conversion events have occurred during the recent human history.

Thus, the human Y chromosome phylogeny has become over time the tool of choice for the analysis of gene conversion at population level (Bosch et al. 2004; Trombetta et al. 2010, 2014, 2016; Hallast et al. 2013; Skov et al. 2017; Shi et al. 2018, 2019) helping to investigate the dynamics of a mechanism that seems to be evolved to protect essential functions of particular genomic regions.

Y chromosome reference sequences

The human reference sequence of Y chromosome

The reference sequence for the human genome is available from the UCSC genome browser (http://genome.ucsc.edu) and was produced by the International Human Genome Sequencing Consortium (2001).

The Y chromosome sequence deposited in the UCSC genome browser is almost entirely based on a single male donor. In total, the sequence covered approximately 23 Mb of the MSY, including both Yq and Yp sequences, and provides finished nucleotide sequence for roughly 97% of the MSY euchromatin (Skaletsky et al. 2003). Three gaps remained in the final assembly, one of which corresponding to the centromere, but recently some attempts have been made to obtain the sequence of the centromeric region of Y (Jain et al. 2018).

The analysis of the biallelic polymorphisms of the Y chromosome revealed that the majority of the Y-chromosomal reference sequence derives from a single chromosome belonging to haplogroup R. However, a 0.8 Mb portion on the Yq, corresponding to the AZFa region, derives from a different man (Sun et al. 1999) and belongs to haplogroup G.

In February 2009 the human reference sequence was updated with the GRCh37/hg19 version (NCBI Build 37.1), produced by the Genome Reference Consortium, while a more recent version available on the UCSC genome browser is that submitted by the same Consortium in 2013, corresponding to GRCh38/hg38 (NCBI Build 38) assembly.

The chimpanzee reference sequence of Y chromosome

Evolutionarily, chimpanzees are the closest living species to humans (Kehrer-Sawatzki and Cooper 2007). Despite the clear phenotypic differences between humans and chimpanzees, at genomic level the two species are very similar each other, with only 1.2% - 1.4% sequence divergence observed (Stone et al. 2002).

Divergence between the human and chimpanzee Y chromosome is known to be approximately 1.7% (Stone et al. 2002), due to the higher mutation rate of the Y chromosome. The human Y chromosome is much larger than the chimpanzee one, covering about 59 Mb, compared to ~35 Mb of the chimpanzee Y (Ross et al. 2005). This is mainly due to the human-specific heterochromatin variable length and many structural differences (Hughes et al. 2010; Hallast and Jobling 2017) implying rapid evolution during the past 6 million years.

The chimpanzee Y chromosome sequence is essential for the investigation of gene conversion dynamics in humans, since it may provide evidence for the ancestral state of human sequences and of possible direction of the conversion activity. Moreover, the study of sequence divergence between species can also give an information of whether the gene conversion may be occurring (Rozen et al. 2003). The publication of the finished chimpanzee reference sequence (Hughes et al. 2010) and the subsequent more detailed releases, allowed the human-chimpanzee comparison studies to be performed. In this study, we will refer to the Chimp Jan. 2018 - Clint_PTRv2/panTro6 Assembly to carry out some comparative analysis of P6, P7 and P8 human-chimpanzee conserved palindromes.

AIMS

The biological importance of sex-linked IRs, independently originated in several taxa, and the establishment of inter-paralogs gene conversion have been related to a strong adaptive significance concerning the necessity to maintain the structural integrity of ampliconic genes involved in male-specific functions. For this reason, it has been proposed that gene conversion evolved as a mechanism to retain the ancestral state of gene sequences: a *de novo* mutation on a palindrome arm is preferentially back mutated to the ancestral state rather than transmitted to the other arm. In addition, the bias towards the ancestral state has been also proposed to explain the lower human-chimpanzee sequence divergence observed in palindrome arms compared to the spacers. However, despite thousands of human Y chromosomes have been sequenced in their unique regions, little is known about within-population sequence diversity and evolutionary dynamics of MSY palindromes.

In this context, in order to investigate features and dynamics of Y-Y gene conversion and to provide more information about the evolutionary meaning of NAGC, we performed sequencing analysis of P6, P7 and P8 palindromes in 157 Y chromosomes. We propose to identify as many as possible unbiased PSVs and gene conversion events, and to investigate them across a reliable SNP-defined phylogeny. In particular, we will try to address the following points:

- To estimate the minimum number of mutational events and gene conversions which shaped the diversity of palindromic sequences;

- To test the hypothesis that Y-Y GC is a mechanism biased towards the retention of the ancestral state of variants in MSY palindromes;

- To shed light on the evolutionary dynamics of palindromic sequences by estimating a palindrome-specific gene conversion and mutation rate;

- To examine the mutational dynamics also within palindrome spacers through the estimate of a spacer-specific mutation rate;

- To estimate the gene conversion tract-length, when possible.

Moreover, since palindrome P8 shows some peculiar features, due to the high similarity (>90%) shared with 4 gametologous sequences on the X chromosome and the presence of the VCX/VCY coding genes, we also aimed to investigate the role of the inter-chromosomal gene conversion in shaping the variability and evolution of P8 palindrome. In particular we propose to:

- Characterise the X-to-Y gene conversion in the whole gametologous region;

- Compare the effects of gene conversion between functionally different regions of P8 palindrome, paying particular attention to the VCY gene;

- Analyse the X-to-Y GC tract-length and the dynamics of the inter-chromosomal gene conversion through the estimate of a X-to-Y gene conversion rate.

RESULTS

Sample selection

The human Y chromosome phylogeny represents the most prominent investigative tool to fully clarify the dynamics of inter-paralog gene conversion (Bosch et al. 2004; Trombetta et al. 2010, 2014, 2016; Hallast et al. 2013; Skov et al. 2017). To this aim, we firstly reconstructed the evolutionary relationships among the 157 Y chromosomes selected to be analysed for P6, P7 and P8 palindromes.

The subjects have been chosen from two different dataset (D'Atanasio et al. 2018; Finocchio et al. 2018) (Table 2) using a biased approach. We selected Y chromosomes on the basis of their genetic affiliation: belonging to the most divergent Y tree lineages and distributed in many different populations. On the contrary, an unbiased sampling method, where subjects are chosen independently from their haplogroup affiliation, could have under-represented the genetic differentiation within the sample, especially that of rare lineages.

This preliminary step has been necessary to maximize the variability among our Y chromosomes and to obtain a reliable unbiased data also for MSY palindromes.

ID	Haplogroup	References
S101	A00-L1086	D'Atanasio et al. (2018)
S102	A0-V48	D'Atanasio et al. (2018)
S103	A0-V48	D'Atanasio et al. (2018)
S104	A1-M31	D'Atanasio et al. (2018)
S105	A2-PN3	D'Atanasio et al. (2018)
S106	A2-PN3	D'Atanasio et al. (2018)
S107	A3-M28	D'Atanasio et al. (2018)

S108	A3-M51	D'Atanasio et al. (2018)
S109	A3-M51	D'Atanasio et al. (2018)
S110	A3-V3663	D'Atanasio et al. (2018)
S111	A3-V3	D'Atanasio et al. (2018)
S112	A3-V3	D'Atanasio et al. (2018)
S113	A3-V3	D'Atanasio et al. (2018)
S114	A3-V3	D'Atanasio et al. (2018)
S115	A3-V3	D'Atanasio et al. (2018)
S116	A3-V3	D'Atanasio et al. (2018)
S118	A3-V317	D'Atanasio et al. (2018)
S119	A3-V317	D'Atanasio et al. (2018)
S120	A3-V6379	D'Atanasio et al. (2018)
S121	A3-V6379	D'Atanasio et al. (2018)
S122	A3-V3663	D'Atanasio et al. (2018)
S123	A3-V67	D'Atanasio et al. (2018)
S124	A3-V6379	D'Atanasio et al. (2018)
S125	A3-V6379	D'Atanasio et al. (2018)
S126	A3-V3663	D'Atanasio et al. (2018)
S127	A3-V3663	D'Atanasio et al. (2018)
S128	A3-M13*	D'Atanasio et al. (2018)
S129	A3-V3663	D'Atanasio et al. (2018)
S130	B1-M236	D'Atanasio et al. (2018)
S131	B1-M146	D'Atanasio et al. (2018)
S132	B2a-M109	D'Atanasio et al. (2018)
S133	B2a-M109	D'Atanasio et al. (2018)
S134	B2b-M112	D'Atanasio et al. (2018)
S135	E-V44	D'Atanasio et al. (2018)
S136	E-V257	D'Atanasio et al. (2018)
S137	E-V259	D'Atanasio et al. (2018)
S138	E-V259	D'Atanasio et al. (2018)
S139	E-V5459	D'Atanasio et al. (2018)
S140	E-V6873	D'Atanasio et al. (2018)
S141	E-V6873	D'Atanasio et al. (2018)
S142	E-V6873	D'Atanasio et al. (2018)
S143	E-V6873	D'Atanasio et al. (2018)

S144	E-V6873	D'Atanasio et al. (2018)
S145	E-V6873	D'Atanasio et al. (2018)
S146	E-V5459	D'Atanasio et al. (2018)
S147	E-V5459	D'Atanasio et al. (2018)
S148	E-V5459	D'Atanasio et al. (2018)
S149	E-V65	D'Atanasio et al. (2018)
S150	E-V65	D'Atanasio et al. (2018)
S151	E-V65	D'Atanasio et al. (2018)
S152	E-V65	D'Atanasio et al. (2018)
S153	E-V65	D'Atanasio et al. (2018)
S154	E-3746	D'Atanasio et al. (2018)
S155	E-V5184	D'Atanasio et al. (2018)
S156	E-V5184	D'Atanasio et al. (2018)
S157	E-V3746	D'Atanasio et al. (2018)
S158	E-V3746	D'Atanasio et al. (2018)
S159	E-V3746	D'Atanasio et al. (2018)
S160	E-V2009	D'Atanasio et al. (2018)
S161	E-V2009	D'Atanasio et al. (2018)
S162	E-V22	D'Atanasio et al. (2018)
S163	E-V5001	D'Atanasio et al. (2018)
S164	E-V5001	D'Atanasio et al. (2018)
S165	E-Z15939	D'Atanasio et al. (2018)
S166	E-V5001	D'Atanasio et al. (2018)
S167	E-A186	D'Atanasio et al. (2018)
S168	E-A186	D'Atanasio et al. (2018)
S169	E-L516	D'Atanasio et al. (2018)
S170	E-Z15939	D'Atanasio et al. (2018)
S171	E-L516	D'Atanasio et al. (2018)
S173	E-Z15939	D'Atanasio et al. (2018)
S175	E-L516	D'Atanasio et al. (2018)
S176	E-V4257	D'Atanasio et al. (2018)
S177	E-L516	D'Atanasio et al. (2018)
S178	C-V20	D'Atanasio et al. (2018)
S179	C-M8	D'Atanasio et al. (2018)
S181	J2a-L26	D'Atanasio et al. (2018)

S182	J1-M267	D'Atanasio et al. (2018)
S183	R-V1589	D'Atanasio et al. (2018)
S184	R-V1589	D'Atanasio et al. (2018)
S185	R-V1589	D'Atanasio et al. (2018)
S186	R-V69	D'Atanasio et al. (2018)
S187	R-V69	D'Atanasio et al. (2018)
S188	R-V69	D'Atanasio et al. (2018)
S189	R-V69	D'Atanasio et al. (2018)
S190	R-V1589	D'Atanasio et al. (2018)
S191	R-V69	D'Atanasio et al. (2018)
S192	R-V1589	D'Atanasio et al. (2018)
S193	R-V1589	D'Atanasio et al. (2018)
S194	R-V1589	D'Atanasio et al. (2018)
S195	R-V8	D'Atanasio et al. (2018)
S196	R-V1589	D'Atanasio et al. (2018)
S197	R-V4453	D'Atanasio et al. (2018)
S198	R-V1589	D'Atanasio et al. (2018)
S200	R-V1589	D'Atanasio et al. (2018)
S201	R-V69	D'Atanasio et al. (2018)
S202	R-V69	D'Atanasio et al. (2018)
S203	R-V1589	D'Atanasio et al. (2018)
S204	R-V5776	D'Atanasio et al. (2018)
S206	J1-M267	D'Atanasio et al. (2018)
S207	J1-M267	D'Atanasio et al. (2018)
S208	J1-M267	D'Atanasio et al. (2018)
S209	J1-M267	D'Atanasio et al. (2018)
S210	J1-M267	D'Atanasio et al. (2018)
S300	J2a-L26	Finocchio et al. (2018); present study
S301	J1-M267	Finocchio et al. (2018); present study
S302	J2a-L397	Finocchio et al. (2018); present study
S304	J2a-M92	Finocchio et al. (2018); present study
S305	J2a-L26	Finocchio et al. (2018); present study
S306	J1-M267	Finocchio et al. (2018); present study
S307	J2b-M12	Finocchio et al. (2018); present study
S308	J2a-M92	Finocchio et al. (2018); present study

S309	J2a-M92	Finocchio et al. (2018); present study
S310	J2a-M410* (×L26)	Finocchio et al. (2018); present study
S311	J2a-L26	Finocchio et al. (2018); present study
S312	J2a-M67	Finocchio et al. (2018); present study
S313	J1-M267	Finocchio et al. (2018); present study
S314	J2a-L26	Finocchio et al. (2018); present study
S315	J2a-M410* (×L26)	Finocchio et al. (2018); present study
S316	J2a-M67	Finocchio et al. (2018); present study
S317	J2b-M12	Finocchio et al. (2018); present study
S318	J2a-M410* (×L26)	Finocchio et al. (2018); present study
S319	J2b-M12	Finocchio et al. (2018); present study
S320	J2a-M92	Finocchio et al. (2018); present study
S321	J2b-M12	Finocchio et al. (2018); present study
S322	J1-M267	Finocchio et al. (2018); present study
S323	J2a-M67	Finocchio et al. (2018); present study
S324	J2a-L26	Finocchio et al. (2018); present study
S325	J2a-L397	Finocchio et al. (2018); present study
S326	J2a-L397	Finocchio et al. (2018); present study
S327	J2a-L397	Finocchio et al. (2018); present study
S328	J2a-M410* (×L26)	Finocchio et al. (2018); present study
S329	J2b-M12	Finocchio et al. (2018); present study
S330	J1-M267	Finocchio et al. (2018); present study
S331	J2a-M67	Finocchio et al. (2018); present study
S332	J2b-M12	Finocchio et al. (2018); present study
S333	J2a-M410* (×L26)	Finocchio et al. (2018); present study
S334	J2b-M12	Finocchio et al. (2018); present study
S335	J2a-L397	Finocchio et al. (2018); present study
S336	J1-M267	Finocchio et al. (2018); present study
S337	J2a-M67	Finocchio et al. (2018); present study
S338	J2a-L26	Finocchio et al. (2018); present study
S339	J2a-M92	Finocchio et al. (2018); present study
S340	J2a-M67	Finocchio et al. (2018); present study
S341	J2a-M410* (×L26)	Finocchio et al. (2018); present study
S342	J1-M267	Finocchio et al. (2018); present study
S343	J2a-L397	Finocchio et al. (2018); present study

S344	J2a-L26	Finocchio et al. (2018); present study
S345	J2a-M67	Finocchio et al. (2018); present study
S346	J2a-M92	Finocchio et al. (2018); present study
S347	J2a-M67	Finocchio et al. (2018); present study
S348	J2a-L397	Finocchio et al. (2018); present study
S349	J2a-L397	Finocchio et al. (2018); present study
S350	J2a-L26	Finocchio et al. (2018); present study
S351	J2a-L26	Finocchio et al. (2018); present study
S352	J2b-M12	Finocchio et al. (2018); present study
S353	J1-M267	Finocchio et al. (2018); present study

Table 2*:* Samples analysed by Next Generation Sequencing for both X-degenerate and palindromic regions (P6, P7 and P8). Haplogroup affiliation is expressed according to the nomenclature “by lineage” and “marker”. The DNA samples are from blood, saliva or cell lines.

MSY phylogeny and time estimate

Phylogenetic tree of Y chromosomes

For the phylogenetic tree reconstruction (Figure 12) and time estimates, we re-analysed the genetic variation within ~3.3 Mb of the X-degenerate portion of the MSY in 157 Y chromosomes that were sequenced at high-depth (about 50×) in two previous studies (D’Atanasio et al. 2018; Finocchio et al. 2018). Moreover, we included in the same analysis 4 precisely radiocarbon-dated ancient specimens (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015), used as calibration points for an accurate time estimate.

Our new SNP calling analysis revealed a total of 7,240 mutational events which occurred in 7,206 positions, whose 9 resulted to be tri-allelic and 23 recurrent. In addition, 57 positions which were invariant in the entire sample set, but different from the reference sequence (GRCh37/hg19), have been interpreted as reference-specific mutations and were not considered for further phylogenetic analysis. In summary, 6,135 out of 7,240 SNPs were already described in D’Atanasio et al. (2018), while 811 positions, not present in the latter study, were found to be variable in Finocchio et al. (2018). Of the remainder variants, 71 additional SNPs resulted to be in common with the ISOGG dataset (September 2020) and 2 with dbSNP 151 (UCSC Genome Browser). Interestingly, we discovered 221 new bi-allelic polymorphisms in the haplogroup J, falling within the blocks of sequences not analysed by Finocchio et al. (2018).

More in general, the topology of our phylogenetic tree (Figure 12) was found to be coherent with some recently published world-wide Y phylogenies (Karmin et al. 2015; Poznik et al. 2016; D’Atanasio et al. 2018). The deepest branch of the unrooted tree (branch 1 in Figure 12) is characterised by 857 mutations (Figure 12) resulting from the comparison with the reference sequence. Since these variants have been found to be different between the A00 sample and all other samples, we cannot exactly define how many mutations are private of the A00 chromosome or occur at the A0-T branch, so we considered these two branches together as branch 1 (Figure 12).

We observed an excess of transitions compared to transversions, resulting in a TiTv ratio = 1.73 (4584 vs 2656), that is coherent with previous findings on the MSY variability (Scozzari et al. 2014; Trombetta et al. 2015b; Helgason et al. 2015) and with the hypothesis of an excess of C to T (or G to A) modifications at 5’-CpG-3’ sites (Kong et al. 2012).

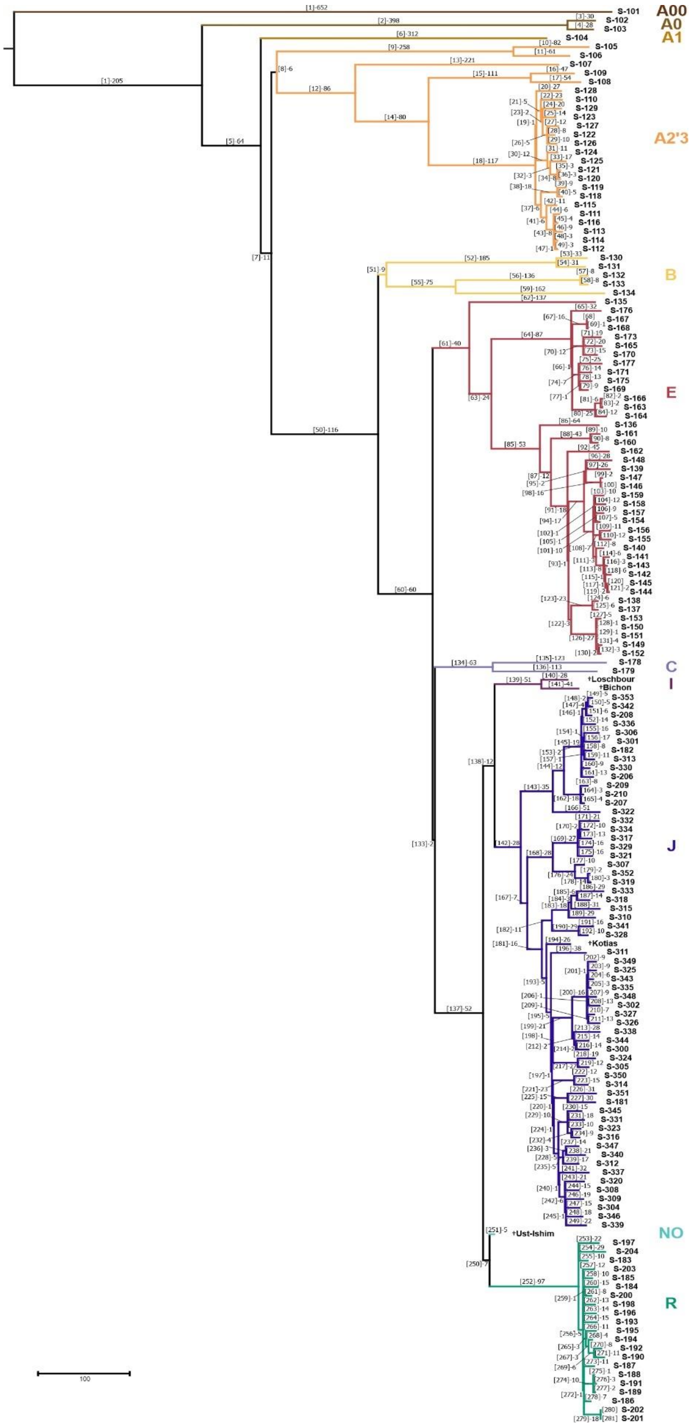
S-101
A00
S-102
S-103
A0
S-104
A1
S-105
S-106
S-107
S-109
S-108
S-128
S-110
S-129
S-123
S-127
S-122
S-126
S-124
S-125
S-121
S-120
S-119
S-118
S-115
S-111
S-116
S-113
S-114
S-112
A2'3
S-130
S-131
S-132
S-133
S-134
B
S-135
S-176
S-167
S-168
S-173
S-165
S-170
S-177
S-171
S-175
S-169
S-166
S-163
S-164
E
S-136
S-161
S-160
S-162
S-148
S-139
S-147
S-146
S-159
S-158
S-157
S-154
S-156
S-155
S-140
S-141
S-143
S-142
S-145
S-144
S-138
S-137
S-153
S-150
S-151
S-149
S-152
S-178
S-179
C
†Loschbour
†Bichon
I
S-353
S-342
S-208
S-336
S-306
S-301
S-182
S-313
S-330
S-206
S-209
S-210
S-207
S-322
S-332
S-334
S-317
S-329
S-321
J
S-307
S-352
S-319
S-333
S-318
S-315
S-310
S-341
S-328
†Kotias
S-311
S-349
S-325
S-343
S-335
S-348
S-302
S-327
S-326
S-338
S-344
S-300
S-324
S-305
S-350
S-314
S-351
S-181
S-345
S-331
S-323
S-316
S-347
S-340
S-312
S-337
S-320
S-308
S-309
S-304
S-346
S-339
NO
†Ust-Ishim
S-197
S-204
S-183
S-203
S-185
S-184
S-200
S-198
S-196
S-193
S-195
S-194
S-192
S-190
S-187
S-188
S-191
S-189
S-186
S-202
S-201
R
100

Figure 12 (previous page): Phylogenetic relationships among the 161 samples analysed. At the tip of each branch the ID sample is reported. The branch nomenclature (in brackets) and the number of mutational events defining each branch is shown above (or near) it. Branch lengths are proportional to the number of mutations. To the right the main haplogroups are indicated.

Mutation rate and dating

The 4 ancient archeologically-dated Y chromosomes (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015) have been used as calibration points for the dating of the tree nodes. We obtained a mutation rate for the X-degenerate region of 7.39×10^{-10} mutations/base/year (sd $\pm 0.38 \times 10^{-10}$), which is coherent with the estimates in other studies (Fu et al. 2014; Trombetta et al. 2015b; D'Atanasio et al. 2018; Finocchio et al. 2018). This figure corresponds to 1 mutation every 406.6 years (sd ± 21 years).

We used this mutation rate to obtain an accurate estimate of the coalescence age of the tree nodes by applying the Rho statistics. We obtained a time to the most recent common ancestor (TMRCA) for the 157 Y chromosomes of 255.4 kya (95% CI 239.1-271.7 kya). This estimate resulted to be close to that reported in other studies based on the A00-T coalescence (Mendez et al. 2013; 2016; Hallast et al. 2015; Karmin et al. 2015; D'Atanasio et al. 2018).

Preliminary analysis on P6, P7 and P8 palindromes

The 157 DNA samples selected for this study were analysed by Next Generation Sequencing, with an average depth of 50×, for P6, P7 and P8 palindromes. We obtained a total of 136,818 bp/each sample, after discarding the interspersed repetitive elements.

Mapping of duplicated reads and tricky variant calling

The standard approaches of NGS are not suitable for ampliconic regions of the genome because of the disproportion between the length of sequencing reads (about 100-200 bp) and the length of amplicon-repeat units, that in the human Y chromosome range from ~10 kb to more than 1 Mb. Generally, each read deriving from different highly similar repeats could be not-univocally mapped within the reference genome, because NGS methods do not incorporate mapping information. For this reason, we carefully considered all possible cases of read 'mis-mapping' (Figure 13).

More specifically, in case of palindromes, reads deriving from the sequencing of one arm will be mapped against both palindrome arms of the reference, producing a double value (2N) of the sequencing depth with respect to the depth (N) of a non-duplicated region, such as the spacer.

A challenging issue is that this read mis-mapping may strongly affect the automatic SNP-calling procedure, especially in the case of a 'pseudo-heterozygous' reference. Indeed, in presence of a reference PSV (for example a T/G – proximal arm/distal arm), depending on the genotype of the sequenced sample, we will have different results (Figure 13). In particular, the reads of a T/G sample resulted to be accurately mapped at the corresponding paralogous sites of the T and G bases of the 'pseudo-heterozygous' reference, so that a N depth value is observed at both sites (Figure 13A). In this case, no real SNPs are present and no SNPs are called by the automatic procedure. Conversely, the reads of the two 'pseudo-homozygous' states, G/G and T/T, are found to be completely mapped against the paralogous position of the reference sequence showing the same base, thus returning a DP value = 2N at this site and DP = 0 at the other paralogous site (Figure 13A). In this case, the automatic SNP calling will fail in the identification of a new SNP (actually present in the site showing the DP = 0), which can be identified only through the analysis of the read distribution over the paralogous sites.

In case of a ‘pseudo-homozygous’ reference sequence (Figure 13B), the NGS reads of the three possible ‘pseudo-diploid’ states (T/G, G/G and T/T) are mapped twice and randomly against both paralogous sites of the reference, returning a DP = 2N for each position (Figure 13B). As a result, the SNP-calling of the ‘pseudo-heterozygous’ sample will return two SNPs (one at each paralogous positions) while only one SNP should be actually present (at proximal or distal position).

In our study, these problems did not affect subsequent analyses if properly taken into account.

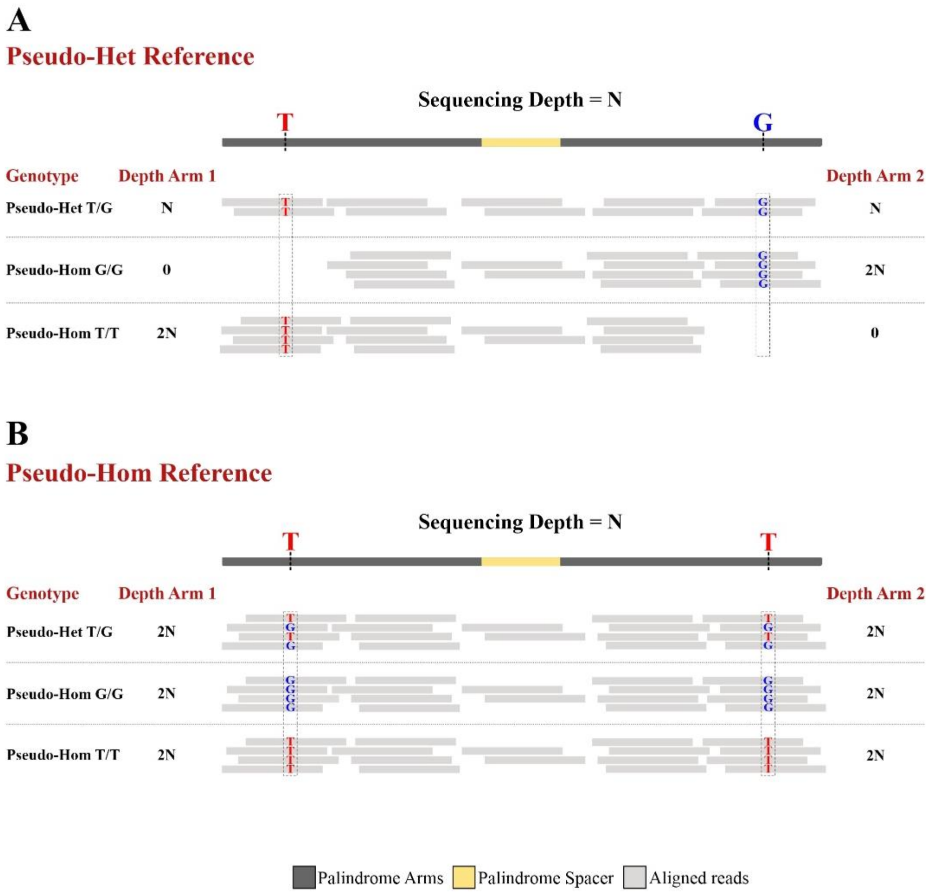

Figure 13: Alignment of NGS reads for three possible ‘pseudo-diploid’ genotypes against a 'pseudo-heterozygous' **(A)** and a 'pseudo-homozygous' **(B)** reference sequence (GRCh37/hg19). For each genotype the DP value for arms is reported.

Depth Analysis: detection of structural variations

Since the MSY is frequently involved in structural rearrangements, it is not excluded that one palindrome arm is lost by deletion and that the apparent ‘pseudo-homozygous’ states are indeed ‘pseudo-hemizygous’ ones. For this reason, we successfully assessed the presence of both arms of each palindrome in all 157 samples by using primers pairs overlapping the sequences between arms and unique regions.

Then, we performed the bioinformatic analysis of depth (see Methods), which returned EMA (Exponential Moving Average) values for each sample showing strong fluctuations along the entire palindrome, possibly due to a differential targeting of regions during the preliminary steps of sequencing. Despite this, we observed in palindrome arms an average EMA value ~2 times higher than the average EMA of the associated spacer, for all the three palindromes here analysed.

In Figure 14, we present an instance of EMA trend in palindrome P7 for S-108 sample. The average EMA calculated within arms is ~1.64, compared to a value of ~0.91 for the spacer, about 1.8 times higher. It is worth noting that, in proximity of the T/G PSV position of the reference sequence (chrY: 17987068/18016494), designated with ‘*’ in the graph, the ‘pseudo-homozygous’ G/G sample S-108 shows EMA ~0 at the proximal site (chrY: 17987068), and EMA ~3 at the distal one (chrY: 18016494), indicating that all paralogous reads are mapped at the distal position of the reference PSV, showing the G nucleotide.

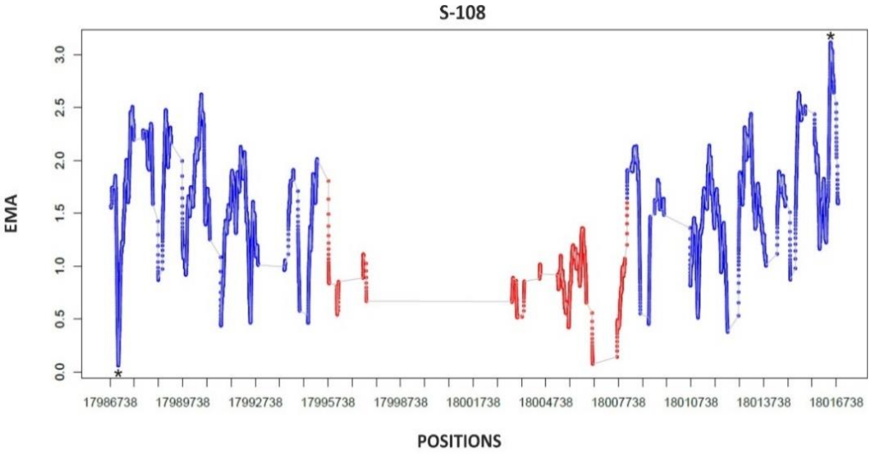

Figure 14: Exponential moving average trend in P7 palindrome for the ’pseudo-homozygous’ sample S-108. On the axes are shown palindrome positions (X) and the associated EMA values (Y) for arms (in blue) and spacer (in red). The interspersed repeated elements (grey lines) have been cut off from the analysis.

The reproducibility of EMA trend in the whole set of samples allowed us to perform a second phase of depth analysis, which consisted in using the depth information to identify duplications or deletions within our target regions. However, due to i) the duplicated nature of palindromes, ii) the possible imbalance in the amplification of the regions introduced by the WGA method (used for 59 out of 157 samples), and iii) the strong oscillations of the depth values, the identification of putative duplications and ‘pseudo-hemizygous’ deletions has been pretty hard to be achieved in paralogs. Therefore, we have dwelt specifically on the ‘pseudo-homozygous’ deletions present in at least 1 sample, and on blocks of sequences that may have been indicative of structural rearrangements (both duplications and deletions) present in ≥ 2 phylogenetically related samples, probably generated by a single recombinative event.

In summary, we found no evidence of duplications and ‘pseudo-hemizygous’ deletions within P6, P7 and P8 palindromes, neither in the arms nor in the spacer. Conversely, we identified a ~1.4 kb deletion in palindrome P6, present in a ‘pseudo-homozygous’ state in two phylogenetically related Y chromosomes, belonging to A2-PN3 lineage (Figure 15).

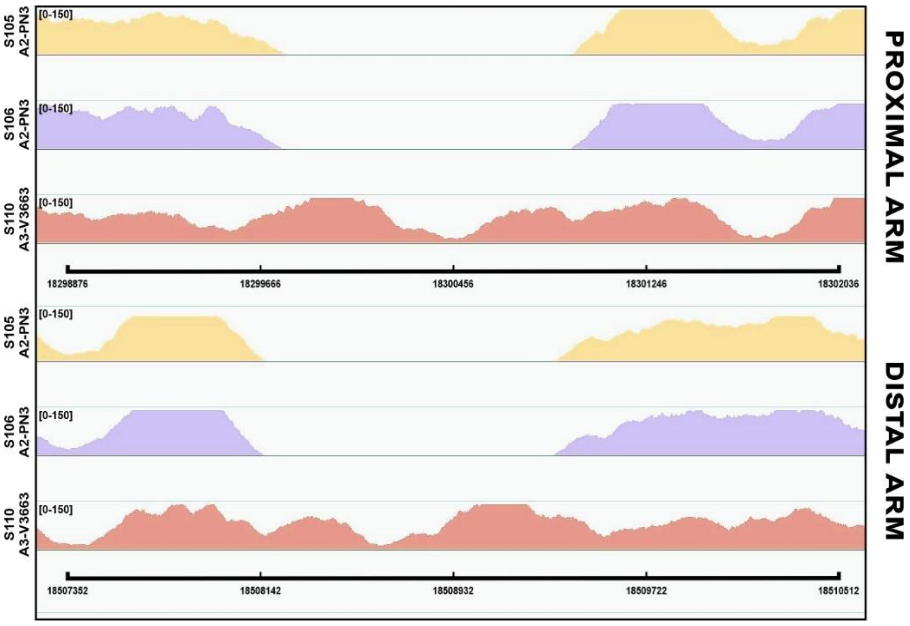

Figure 15: Coverage profile of the two A2-PN3 deleted samples (S105-S106) and of a non-deleted control (S110) for both proximal and distal arms of P6 palindrome. The range of variability of depth values (in squared brackets) has been set to 0-150.

The deletion has been experimentally validated through PCR with two primer couples: 1) FOR: 5’–TCTTGTGGCCTCTGGCTACT–3’ and REV 5’–AAAACCAGTTTATTGAAGTATGGTTGT–3’ and 2) FOR 5’–TTGCCATTTGGGTTTTGATT–3’ and REV 5’–GTGCCCAAGATGTCCGTTAC–3’, which fall inside and outside the deletion, respectively. We also tested the putative deletion in an additional A2-PN3 chromosome, for a total of three A2-PN3 samples and three controls (Figure 16), belonging to different haplogroups. In the first validation, we obtained a 222 bp amplicon only for the three positive non-deleted controls, as expected (Figure 16A). In the second case, we selected primers outside the deletion that should have amplified 1910 bp in the reference genome (in silico PCR). We obtained a ~600 bp deleted amplicon for the three A2-PN3 samples, but no amplification has been obtained for the controls, since the PCR protocol failed in generating such a long product (Figure 16B).

These findings suggest that the deletion arose specifically in the A2-PN3 lineage.

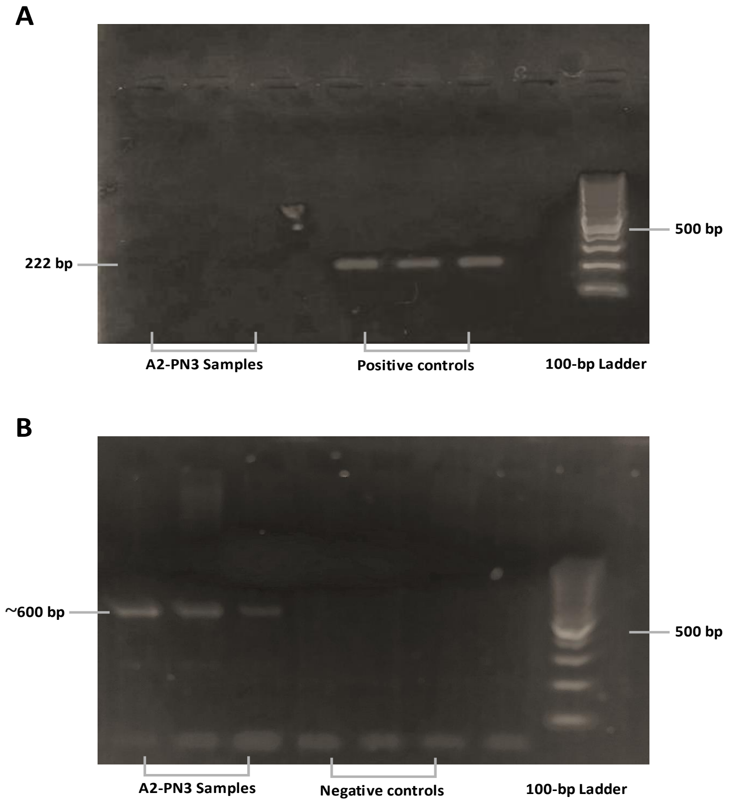

Figure 16: PCR products obtained with two different in-house primer pairs, used for the amplification of three A2-PN3 subjects and three controls belonging

to different haplogroups. **A)** Amplification provided with primers located inside the deletion. The PCR product was obtained only for the controls. **B)** Complementary amplification with primers located outside the deletion. In this case we observed a shorter ~600 bp PCR product from A2-PN3 samples, compared with that expected from controls, based on in silico PCR.

From the sequencing of PCR products of the A2-PN3 samples we identified a deletion of 1,393 bp on both proximal (chrY:18299763-18301155) and distal (chrY:18507948-18509340) arms of P6 palindrome. Moreover, by sequence analysis, we found two identical 217 bp direct repeats (DRs), respectively upstream and downstream the deleted fragment.

From these data, we hypothesized that an intra-chromosomal homologous recombination (HR) between the two DRs of one arm has occurred in the A2-PN3 lineage (branch 9 – Figure 12), generating the deletion which has been subsequently transferred by a gene conversion event on the other arm of P6.

Genetic variability at three palindromic sequences

From the NGS analysis, we obtained sequences for a total of 136,818 bp of P6, P7 and P8 palindromes (Table 3) after the removal of interspersed repeated elements. Thanks to the accurate variant calling and filtering criteria adopted (see Methods), each genotype was recorded as ‘pseudo-heterozygous’ or ‘pseudo-homozygous’. Following the maximum parsimony principle, we identified a total of 206 PSVs and 136 gene conversion events within the three targeted palindromes. In addition, 61 mutations have been detected within the three non-duplicated spacers.

Due to the mapping issues described above, we could not assign the phase of most of the ‘pseudo-heterozygous’ variants (Figure 13), except for those mapping at boundaries between duplicates and haploid regions. However, this aspect did not affect the possibility of identifying ‘pseudo-heterozygous’ genotypes and gene conversion events.

Pal	Proximal arm (bp)[a]	Distal arm (bp)[a]	Spacer (bp)[a]	Proximal arm sequenced bp	Distal arm sequenced bp	Spacer sequenced bp
P6	109995	110022	46229	35326	34986	18911
P7	8723	8726	12638	4662	4690	1680
P8	38006[b] 35160	37404[c] 35159	3414	17456 [b] 15225	17221[c] 15230	1886

Table 3: Specifications of the three palindromes analysed in the present study. P6, P7 and P8 are located on the long arm of the Y chromosome; conventionally, the arm of the palindrome closest to the centromere is indicated as ‘proximal’, as a consequence the farthest arm is designated as ‘distal’.

[a] According to Human Feb. 2009 - GRCh37/hg19 Assembly.

[b, c] Palindrome P8 length adding ~2.8 ([b]) and ~2.2 ([c]) kb highly similar flanking regions analysed in the present study.

Genetic diversity at P6 palindrome

P6 is the longest singleton palindrome of the MSY and covers a total of 266 kb on the Yq. Its arms, which share sequence identity of 99.97% in the reference genome, are 110 kb long and are separated by a 46 kb non-duplicated spacer (Skaletsky et al. 2003). The arm-to-arm alignment of the P6 reference sequence revealed 31 single nucleotide-PSVs (SN-PSVs). After eliminating the interspersed repeated elements, we obtained sequencing data for a total of 70,312 bp in the arms (35,326 and 34,986 bp for proximal and distal arm, respectively).

By exploiting the Y evolutionary relationships among our samples, obtained from non-duplicated sequences, we carefully mapped all the mutational events found in P6 across the phylogeny, following the maximum parsimony principle. Thus, we rejected the possibility of back- or new mutations in the same (or paralogous) site (Rozen et al. 2003; Jobling et al. 2013), except for two PSVs (V539 and V570) which resulted to be recurrent (Additional File 1: https://drive.google.com/file/d/1xDtYReofw-furYJdl1gxobExcfZBF5QT/view?usp=sharing).

Our high-resolution analyses disclosed a total of 118 PSVs, 4 of which were already present in reference sequence (V520, V521, V526 and V625) (Additional File 1, Additional Figure 1: https://drive.google.com/drive/folders/1GTnhTR_p1qQ2zUCi9AM_iuO06zGs076E?usp=sharing). In particular, V520 was found to be monomorphic (in a 'pseudo-homozygous' ancestral state) in all the sequenced samples, suggesting that it is a recently arisen PSV of the reference. Except for the reference PSVs, it has not been possible to assess the phase of the arm-to-arm sequence variants because of mapping issues. All the remaining PSVs resulted to be polymorphic in our Y phylogeny. Three PSVs (V540, V586 and V587) showed a peculiar mutational pattern, resulting in a 'pseudo-heterozygous state' or in a converted state in all the samples of the phylogeny, suggesting that a mutation occurred on the stem lineage of the tree, before the human Y chromosome radiation. Considering altogether recurrent mutational events in PSVs V539 and V570, the PSV originated by a mutation arisen in the reference sequence (V520) and mutations generating PSVs before Y chromosome radiation (V540, V586 and V587), the observed diversity of P6 arms can be explained by 116 mutational events occurred within our phylogeny (Additional File 1, Additional Figure 1). The variants appear to be uniformly distributed along the arms and none of them were already described in the database of common SNPs (build 151).

Within the phylogeny, we identified 35 converting sites, 16 of which show multiple gene conversion events. We found in P6 a total of 80 gene conversions, 34 of which restored the ancestral ‘pseudo-homozygous’ state and 45 events occurred towards the fixation of the derived state of variants. For a single event, within V623 PSV, it has not been possible to assess the direction of recombination, since both paralogous positions are not conserved in the orthologous chimpanzee sequence. The highest number of conversion events, leading to the derived ‘pseudo-homozygous’ state (although not statistically significant, $p = 0.2159$, Chi-square test), seems to contradict the hypothesis that Y-Y gene conversion is a molecular mechanism evolved to retain the ancestral state of sequences (Hallast et al. 2013).

Since it is not possible to detect the to-ancestral events occurring on the same branch where the mutation occurred, the total amount of conversions towards the ancestral state could be an underestimate of the true number. Thus, we decided to remove from the count all to-derived gene conversion events that we would not have observed in the phylogeny if they had occurred towards the ancestral. By this approach, we discarded a total of 20 conversions, so the amount of to-derived gene conversions decreased to 25, against the 34 ancestral ones previously mentioned. Also after this calibration, we did not observe a specific bias in the gene conversion mechanism ($p = 0.2413$, Chi-square test).

From our data, we also analysed the GC-biased gene conversion: the tendency towards the fixation of G/C over A/T nucleotide in a gene conversion event. Of the 80 converted PSVs, 3 resulted to be uninformative, since two of them do not alter the GC content (V579 and V619) and for V623 we could not assess the direction of conversion. Among the 77 informative cases (A/G, T/G, A/C and T/C PSVs), 58 resulted in the fixation of GC and 19 of AT nucleotides ($p = 8.8 \times 10^{-6}$, Chi-square test), suggesting a strong GC-bias within P6 palindrome. The existence of the GC-biased gene conversion raises the hypothesis that a bias towards the ancestral state may actually exist, but it can be masked by the GC bias itself. It can happen when, for example, there is a greater number of events in which the derived base is represented by a G or C nucleotide. To test this hypothesis, we perform a new ancestral/derived biased analysis only on the 58 events towards GC bases and we found a higher number of to-derived conversions (34) that is statistically indistinguishable from the to-ancestral ones (24) (p = 0.1892; Chi-square test). We finally discarded from the 34 to-derived events those that we would not have observed if had occurred towards the ancestral state (14), obtaining a final number of 20 to-derived gene conversions and 24 to-ancestral ones (p = 0.5465; Chi-square test). From these results, we can confirm the absence of an ancestral/derived gene conversion bias in P6 palindrome.

From the sequencing of P6 spacer (46,229 bp) we obtained a total of 18,911 bp after removing the repeated sequences. Our variant calling revealed a total of 52 variants across the phylogeny (Table 4), one of which has been found to be a private mutation of the reference sequence (chrY: 18394634) and only one is shared with dbSNP 151 (chrY: 18390543).

POS (GRCh37/hg19)	REF	ALT	HAPLOGROUP	dbSNP 151
18383857	T	G	J2a	
18384669	C	G	A3b1	
18384838	G	A	E-V257/M78	
18389678	T	C	C-M8	
18389752	T	C	R-V88/V69	
18390321	C	T	J2a-M67	
18390543	A	T	E-M2	rs16980497
18390609	G	A	E-M2/Z15939	
18390900	G	A	A3b	
18391299	A	G	J1-M267	
18391323	C	G	A1a-M13	
18391511	G	C	J2b-M12	

18392551	C	A	B2b-M112	
18392789	A	G	A00	
18393473	C	T	J2a-L397	
18393810	A	G	B1-M236	
18393820	G	A	A1a-M31	
18394594	T	C	C-RPS4Y/V20	
18394634	C	T	Reference	
18395243	C	T	A3	
18395249	C	A	E-M2/V5001	
18395639	C	T	A-M13/V317	
18396142	A	T	E-M2	
18397568	G	A	A0	
18400384	C	T	J1-M267	
18400397	A	G	B2b-M112	
18401838	A	G	E-V68/V2009	
18402026	G	C	A1/R-V88	
18403636	T	C	A0/R-V88	
18403739	G	T	J2b-M12	
18403925	A	G	A2-PN3	
18403999	C	T	J2a-M92	
18404545	A	G	E-V257	
18404758	A	C	E-M78/V65	
18412670	T	C	E-M2/E-Z15939	
18412840	C	T	A3	
18413244	A	G	J1-M267	
18413773	T	C	A00	
18415001	A	C	J2a	
18415247	C	T	J2a-L26	
18415639	C	T	A0	
18416443	G	T	A3a	
18417342	G	T	J2a-L397	
18418036	A	G	A0	
18418203	T	A	E-M2	
18418700	G	C	A-M13/V3663	
18419547	A	G	E-V257	
18419580	C	T	B1	
18422249	T	C	J2a-L397	
18423298	A	C	J2b-M12	
18423460	A	G	E-M78/V259	
18425203	C	T	E-M78/V5459	

Table 4: Positions of the 52 mutations identified within palindrome P6 spacer

Genetic diversity at P7 palindrome

P7 is the shortest palindrome of the human MSY, being characterised by a combined length of ~30 kb (8.7 kb each arm) (Table 2). From the arm-to-arm alignment of the reference sequence we found 3 SN-PSVs, however only 1 PSV, located within the non-repetitive regions, has been targeted during the preliminary steps of NGS. In total, we analysed by deep sequencing 9,352 bp in the arms (4,662 and 4,690 for proximal and distal arm, respectively) and 1,680 bp in the spacer (Table 2).

By using the same approach of maximum parsimony principle described above, we identified 16 PSVs showing both 'pseudo-heterozygous' and 'pseudo-homozygous' states (Figure 17), 10 of which have been phased by an arm specific typing (Additional File 2:https://drive.google.com/file/d/1H5HiLQpXdTpilVjj_X0T7fo3iKfWRHlD/view?usp=sharing).

Within the phylogeny, we found a total of 10 gene conversion events, 7 of which occur at V640 (Figure 17, Additional File 2), the only PSV shared with the reference sequence. We could not assess the direction of the V638 PSV since both the deepest clades of the phylogeny and the chimpanzee Y chromosome (Clint_PTRv2/panTro6) share the 'pseudo-heterozygous' T/A genotype at this site, indicating that V638 probably originated before the human-chimpanzee speciation. For this reason, the diversity of P7 palindrome arms is explained by 15 mutational events (Additional File 2).

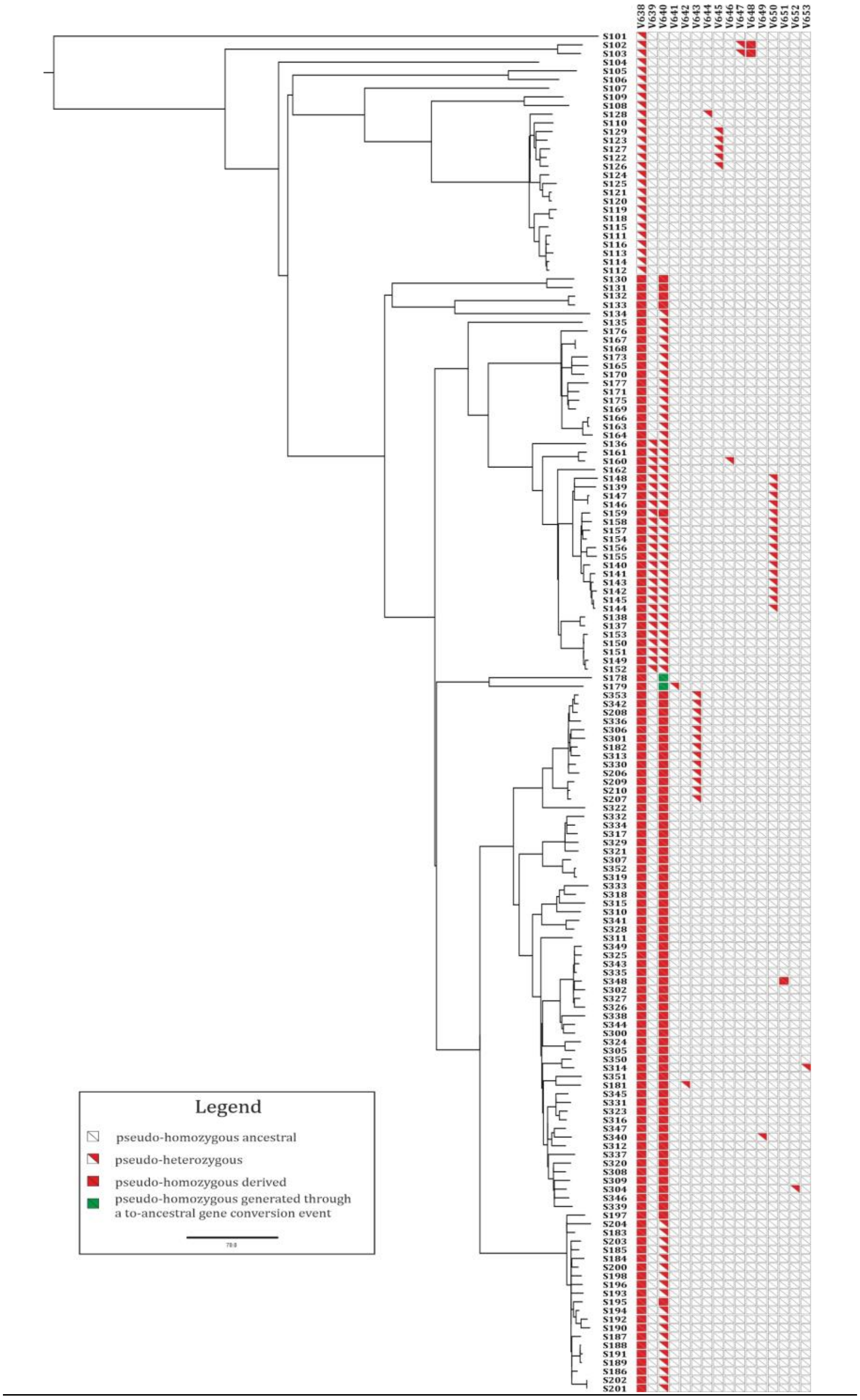
V638
V639
V640
V641
V642
V643
V644
V645
V646
V647
V648
V649
V650
V651
V652
V653
S101
S102
S103
S104
S105
S106
S107
S109
S108
S128
S110
S129
S123
S127
S122
S126
S124
S125
S121
S120
S119
S118
S115
S111
S116
S113
S114
S112
S130
S131
S132
S133
S134
S135
S176
S167
S168
S173
S165
S170
S177
S171
S175
S169
S166
S163
S164
S136
S161
S160
S162
S148
S139
S147
S146
S159
S158
S157
S154
S156
S155
S140
S141
S143
S142
S145
S144
S138
S137
S153
S150
S151
S149
S152
S178
S179
S353
S342
S208
S336
S306
S301
S182
S313
S330
S206
S209
S210
S207
S322
S332
S334
S317
S329
S321
S307
S352
S319
S333
S318
S315
S310
S341
S328
S311
S349
S325
S343
S335
S348
S302
S327
S326
S338
S344
S300
S324
S305
S350
S314
S351
S181
S345
S331
S323
S316
S347
S340
S312
S337
S320
S308
S309
S304
S346
S339
S197
S204
S183
S203
S185
S184
S200
S198
S196
S193
S195
S194
S192
S190
S187
S188
S191
S189
S186
S202
S201
Legend
pseudo-homozygous ancestral
pseudo-heterozygous
pseudo-homozygous derived
pseudo-homozygous generated through a to-ancestral gene conversion event
70.0

Figure 17 (previous page): ‘Pseudo-diploid’ positions identified in the palindrome P7 arms (chrY: 17986738-17995460 and chrY: 18008099-18016824, GRCh37/hg19) by sequencing 157 Y chromosomes. To the left, it is reported the Y chromosome tree showing the phylogenetically relationships of the chromosomes analysed. SNP names are given at the top. Each square is divided into two triangles, representing the paralogous sites in the two arms of the palindrome.

Among the Y-Y gene conversions, we found a significant higher number of events fixing the derived state of mutations, with only 1 case restoring the ancestral 'pseudo-homozygous' genotype (8 vs 1, p = 0.019, Chi-square test). By using the same approach described for P6 palindrome, we excluded the to-derived events not observed in the phylogeny if occurring towards the ancestral state, and we obtained a decreased number of to-derived conversions equal to 6 (p = 0.059, Chi-square test), suggesting, also for P7 palindrome, the absence of a specific trend of the gene conversion mechanism.

Differently from P6 palindrome, we did not observe in P7 the GC-biased gene conversion (6 A/T vs 3 G/C, p = 0.3173, Chi-square test), implying no tendency to fix specific nucleotides.

Within P7 spacer, only 1,680 bp out of the 12,638 bp of its length were sequenced, since represented by non-repeated sequences. From its analysis, we identified 5 mutations across the phylogeny (Table 5).

POS (GRCh37/hg19)	REF	ALT	HAPLOGROUP
17997191	A	G	A3b1-M51
17997207	G	C	J2b-M12
18005289	T	C	J2a-L26
18006025	C	A	C-M8
18007789	A	T	A1a-M31

Table 5: Positions of the 5 mutations identified within 1,680 bp of palindrome P7 spacer.

Genetic diversity at P8 palindrome

The palindrome P8 is made up of two arms each 36 kb long and a 3.4 kb spacer (Table 2). The arms show the highest similarity (99.997%) observed among MSY palindromes (Skaletsky et al. 2003), but we also extended our analysis to two paralogs flanking the arms, long ~2.8 (proximal arm) and ~2.2 kb (distal arm), respectively. These two additional sequences exhibit a lower similarity in the reference genome (91.3%, BLAT result of UCSC) compared to palindrome arms, possibly due to a ~0.6 kb fragment present exclusively in the proximal portion of the palindrome (Figure 18). Moreover, P8 hosts 2 copies of the VCY gene showing 100% identity (Bhowmick et al. 2006), which are expressed uniquely in the testes (Lahn and Page 2000), and share high sequence identity (>90%) with 4 regions on the X chromosome (from ~10 to ~16 kb), each containing a VCX gene copy.

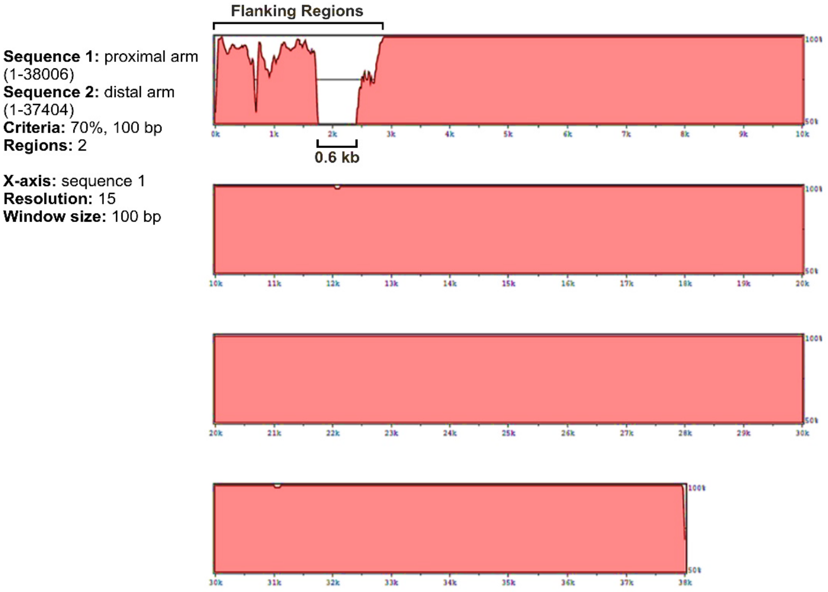

Figure 18: Graphical representation of the arm-to-arm alignment of P8 palindrome, performed with VISTA LAGAN (Brudno et al. 2003). In the figure

are indicated the flanking regions (including the ~0.6 kb difference), which exhibit an average sequence similarity of 91.3% (BLAT result). On the X-axis are shown the proximal arm positions by 1 kb-windows, on the Y-axis is indicated the 50-100% range of sequence identity.

From NGS we obtained 17,456 and 17,221 bp for P8 proximal and distal arm, respectively, and 1,886 bp within the spacer (Table 2). From the analysis of the sequenced regions, we detected 72 polymorphic PSVs across the phylogeny, with the first 16 positions (from V654 to V669) located along the additional flanking regions (Additional File 3: https://drive.google.com/file/d/1SHOusdl4SKje0cS6537-zW0s6vEsSuVM/view?usp=sharing, Additional Figure 2: https://drive.google.com/file/d/17YWRVdl9H8mBX1Q5hPEI9yL8LrC_xksx/view?usp=sharing). Interestingly, the distribution of PSVs along P8 arms is not uniform, with 45 out of 72 PSVs (from V654 to V116*) more densely distributed along the portion showing high similarity with the X chromosome (X-Y gametologous region), compared to the non-similar one (p = 0.018, Fisher Exact test). The observed discrepancy is probably due to the effect of gene conversion from the X chromosome, which helps to increase the Y chromosome variability. Indeed, considering altogether recurrent mutations, mutations introduced by X-to-Y gene conversion and mutations that arose before human Y chromosome radiation, the observed diversity of P8 arms can be explained by 79 mutational events occurred within our phylogeny, more than the number of events generating Y-Y PSVs (Additional File 3, Additional Figure 2).

In our phylogeny, we identified 46 Y-Y gene conversion events within 17 converting PSVs, 8 of which showing frequent conversion activity (Additional File 3, Additional Figure 2). Conversely, the low-similar flanking regions of P8 exhibited evidence of a poor gene conversion: we found a single to-ancestral event within V657 variant, that arose before human-chimpanzee divergence (as well as V659 position). Indeed, it has been suggested that the gene conversion mechanism requires >92% similarity between interacting sequences (Chen et al. 2007), thus the low rate of conversion observed in this region is possibly due to the low inter-paralogue sequence identity represented by the numerous arm-to-arm differences and in particular by the ~0.6 kb structural difference between proximal and distal portions.

Since this ~0.6 kb SV was known to be present only in the reference sequence, we also tested its presence, by PCRs, in other haplogroups of our phylogeny. In particular, we selected subjects belonging to A00, E-M78 and R-V88 haplogroups and we found that all of them host the ~0.6 kb SV between proximal and distal flanking regions of P8. In addition, the BLAT analysis between human and chimpanzee orthologs, revealed that chimpanzee P8 palindrome lacks the ~0.6 kb SV, suggesting that it possibly appeared in the human Y lineage after species separation.

In order to assess the direction of the mutational events (and so of gene conversion) occurring within PSVs that show 'pseudo-hemizygosity' in P8 chimpanzee sequence (V695, V697, V704, V711), in agreement with a more parsimonious hypothesis, we considered the ancestral deleted base to be invariant with respect to its paralogous base. Then, by analysing the direction of gene conversion, we observed a non-significant excess of to-derived events (27) compared to the ancestral ones (19) ($p = 0.24$, Chi-square test). However, by calibrating such events through the removal of not-detectable to-derived conversions, as described for the other two palindromes, no bias in the direction of the Y-Y gene conversion has been observed (13 vs 7, $p = 0.1797$, Chi-square test).

On the contrary, we found clear evidence for the GC-biased gene conversion in palindrome P8, favouring the paralogue holding the G or C base among the 36 informative PSVs (29 GC vs 7 AT, $p = 2.4 \times 10^{-4}$). In order to exclude the possibility of an ancestral bias masked by the GC-nucleotide fixation, we only tested the 29 GC events, which revealed no differences between ancestral (13) and derived (16) conversions (p = 0.5775). From the GC-derived conversions, we then discarded 8 events that we could not have detected if occurred towards the ancestral, obtaining 13 ancestral events vs 8 derived ones (p = 0.2752). This result confirmed the absence of a significant ancestral/derived bias, making the GC-bias the unique apparent driving force of the Y-Y non-allelic recombination.

From the analysis of the 1,886 out of 3,414 bp sequenced in P8 spacer, we detected only 5 variants in our phylogeny. These positions are listed in Table 6.

POS (GRCh37/hg19)	REF	ALT	HAPLOGROUP
16131945	G	C	E/R-V88
16132077	C	T	A3
16132488	G	T	A3b1-M51
16133768	T	A	A-M13/V3663
16134910	G	A	R-V88/V1589

Table 6: Positions of the 5 mutations identified within palindrome P8 spacer.

Mutational pattern of palindromic sequences

Since the palindromic sequences of the Y chromosome behave differently from the rest of Y, showing abundant non-allelic gene conversion events, this increases the possibility of observing homoplasy within phylogeny: i.e. mutations that occur more than once in the same position of the genome, in different directions or on different branches of the phylogeny. In order to evaluate the extent of the recurrence of mutations, we introduced the Recurrence Index (RI), which describes the frequency of mutational events (i.e. every event that introduces sequence changes) within palindrome arms.

We reported for P6, P7 and P8 palindromes RI (%) values of 26.5%, 24.0% and 32.0%, respectively. These values represent the percentage of mutational events occurring in positions that already had 1 mutation. P8 palindrome exhibits the highest, despite not significant value among them, that may be explained by the contribution of the X-to-Y gene conversion (see below). We also calculated the RI for the ~3.3 Mb of the X-degenerate region used to reconstruct the phylogeny, and we obtained a much lower estimate of 0.047%.

By comparing the extent of recurrence between palindrome arms and X-degenerate region, we found that P6, P7 and P8 show a significant higher number of recurrent mutations than the unique region of Y ($p < 0.01$, Fisher Exact test), suggesting that the gene conversion process has a great relevance in increasing homoplasy within palindromes.

Evolutionary dynamics of P6, P7 and P8 palindromes

Estimate of a palindrome-specific gene conversion rate

By mapping gene conversion events within our stable phylogeny, for which we previously estimated the average time of each branch, we obtained the minimum number of independent events that shaped the diversity of palindromes here analysed. We used this value to estimate a minimum and maximum palindrome-specific gene conversion rate with a method reported in this study for the first time.

We obtained for palindrome P6 a rate of 6.01×10^{-6} events per base per year, which ranges between a minimum value of 4.42×10^{-6} and a maximum value of 9.38×10^{-6}; whereas for P7 palindrome we estimated a rate of 3.94×10^{-6} events per base per year, which varies from 3.40×10^{-6} to 4.69×10^{-6}.

While these two values are statistically indistinguishable (p = 0.2057, Comparison of two rates test) a lower conversion rate has been observed for P8 palindrome, corresponding to 2.09 × 10^{-6} (1.98-2.20 × 10^{-6}) events per base per year. This value is significantly lower than P6 gene conversion rate ($p < 0.0001$, Comparison of two rates test), but no different (though slightly) from that of P7 palindrome (p = 0.0639).

In order to better investigate the observed discrepancy, we estimated two distinct gene conversion rates, respectively for P8 X-Y-gametologous region and the non-gametologous one. Interestingly, we found two highly divergent rates, with the former corresponding to 8.7 × 10^{-7} (8.38-9.05 × 10^{-7}) events per base per year, that resulted to be approximately 24 times lower than gene conversion rate showed by the non-gametologous portion, equal to 2.08 × 10^{-5} (1.97-2.18 × 10^{-5}) events per base per year ($p < 0.0001$).

Our results represent an underestimate of the true gene conversion rates since we cannot identify all possible events towards the ‘pseudo-homozygous’ ancestral state, because of the impossibility to detect such events occurring on the same branch where the mutation took place. Despite this, our findings point that:

1. Palindrome P6 exhibits the highest gene conversion rate among the three palindromes here analysed, showing that each base is involved in a Y-Y gene conversion event 6 times every 1 million years, on average. Considering a 25-years human generation, this corresponds to a rate of 1.5 × 10^{-4} conversions per base per generation. Thus, in the transition from father to son, we expect to observe an average of about 16 bases interested by gene conversion within the 110 kb of the palindrome arm.

2. Highly different gene conversion dynamics emerge from two different regions of palindrome P8. For a 25-years generation time, we expect about 4 bases interested by gene conversion within the X-Y-similar region (~16 kb) and more than 11 bases within the non-gametologous region (~22 kb).

Mutation rate of P6, P7 and P8 palindrome arms and spacers

In order to obtain a palindrome-specific mutation rate, we considered the minimum number of independent mutational events occurring during the whole phylogenetic time.

We obtained for P6 and P7 palindromes close estimates of 5.6 × 10^{-10} (sd ± 0.29 × 10^{-10}) and 5.45 × 10^{-10} (sd ± 0.28 × 10^{-10}) mutations/base/year, respectively. Both rates are statistically different from the pedigree-based mutation rate calculated for all MSY palindromes by Helgason et al. (2015), corresponding to 7.37 × 10^{-10} (CI: 6.41–8.48 × 10^{-10}) mutations per position per year (p = 0.018; Comparison of two rates test).

Interestingly, a slightly higher rate of 7.74 × 10^{-10} (sd ± 0.4 × 10^{-10}) mutations/base/year has been found in P8, suggesting an increased tendency to accumulate mutations for this palindrome. Similarly to gene conversion rate, we performed two distinct estimates of P8 mutation rate, respectively for X-Y gametologous region and the non-gametologous one. We found a much higher mutation rate, corresponding to 1.05 × 10^{-9} (sd ± 0.05 × 10^{-9}) mutations/base/year, in the X-Y region compared to the non-gametologous segment, where we observe a rate of 5.11 × 10^{-10} (sd ± 0.26 × 10^{-10}) mutations/base/year (p = 0.0018).

This result highlights different dynamics in the evolution of distinct portions of palindrome P8, since the non-gametologous region behaves exactly like P6 and P7 palindromes, exhibiting indistinguishable mutation rates (p = 0.67 and p = 0.84, respectively), suggesting that its dynamics is in line with that of other MSY palindromes. On the contrary, the gametologous portion of P8, characterised by the highest mutation rate and the lowest rate of gene conversion observed, seems to follow an autonomous evolutionary history.

It should be noted that our calculation represents an underestimate of the actual mutation rate because it does not consider the mutations which generate new PSVs immediately converted to the ancestral state by gene conversion. So, we also performed a new estimate of the mutation rate by incorporating mutations which have been converted to the ‘pseudo-homozygous’ derived state and that we would not have observed if conversion had occurred towards the ancestral. Our consideration was based on the observation that we found no conversion bias towards the ancestral state, so the number of mutations which are immediately converted to the ‘pseudo-homozygous’ derived state should be equal to the number of mutations converted to the ancestral, which are invisible through sequencing. By this approach, we calculated new mutation rates of 6.18×10^{-10} (sd $\pm 0.32 \times 10^{-10}$) and 6.17×10^{-10} (sd $\pm 0.32 \times 10^{-10}$) mutations/base/year for P6 and P7 palindromes respectively, that resulted to be indistinguishable, whereas the new increased estimate for P8 arms corresponds to 8.62×10^{-10} (sd $\pm 0.45 \times 10^{-10}$) mutations/base/year. More in details, within P8 we found a statistically higher mutation rate ($p = 0.004$) in the X-Y gametologous region ($1.13 \times 10^{-9} \pm 0.06 \times 10^{-9}$ mutations per base per year) than in the non-gametologous one ($6.06 \times 10^{-10} \pm 0.31 \times 10^{-10}$). It is worth noting that the mutation rate of the non-gametologous region of P8 (i.e. exhibiting only Y-Y identity) is perfectly in line with mutation rates of P6 and P7 palindromes ($p = 0.9194$ and $p = 0.9508$, respectively).

By estimating the mutation rate of the three haploid spacers, we found an average rate of 9.16×10^{-10} (sd $\pm 0.47 \times 10^{-10}$) and 1.01×10^{-9} (sd $\pm 0.052 \times 10^{-9}$) mutations/base/year for P6 and P7 spacer, respectively. In particular, P6 spacer mutation rate (but not the mutation rate of P7 spacer, for which we only had 1,680 sequenced bp) resulted to be significantly higher than both mutation rates estimated within arms ($p = 0.0031$ and $p = 0.0169$). This difference may explain the higher human-chimpanzee divergence between spacers with respect to the arms, previously suggested by Rozen et al. (2003) for some palindromes of the MSY, in absence of the to-ancestral biased gene conversion within arms restoring the non-mutated state of variants.

Moreover, P8 spacer mutation rate, corresponding to 9.0×10^{-10} mutations/base/year (sd $\pm 0.47 \times 10^{-10}$), was found to be coherent with the rate of P6 and P7 spacer, but also statistically indistinguishable from both mutation rates calculated within the full-length arms of P8 ($p = 0.7421$ and $p = 0.9242$). When comparing P8 spacer mutation rate with both rates of the non-gametologous region of P8, it is found not significantly higher than both, despite these are in line with mutation rates of P6 and P7 arms. Similarly to P7 palindrome, this is possibly due to the low statistical power consequent to the low number of sequenced bases in P8 spacer.

However, our results about P8 palindrome remark the contribution of the X-to-Y gene conversion in increasing variability and mutation rate within P8 arms, together with a weak rate of gene conversion that erases mutations in this region.

Y-Y gene conversion tract-length

Since it is not feasible to determine the exact length of a non-allelic gene conversion tract, a minimum and maximum length is usually estimated. The minimum estimated length corresponds to the distance between the outermost converted PSVs, although it represents an underestimate of the actual value. It is not possible to determine precisely where the conversion broke outside this segment, since non-variant sites are not informative for this purpose. Therefore, the maximum length of the gene conversion segment is defined as the distance between the two nearest non-converted PSVs flanking the converted sites.

From our data, we only identified in P6 palindrome a possible minimum gene conversion tract of 9,011 bp length in the A00-S101 sample. This segment, which involves 4 PSVs of P6 (V617, V619, V623 and V626), is longer than the ectopic gene conversion tract-length usually observed in mammals (Zangenberg et al. 1995; Papadakis and Patrinos 1999; Bosch et al. 2004; Hallast et al. 2005; Bagnall et al. 2005), but coherent with that already observed by Hallast and colleagues (2013) in the same palindrome. The maximum length estimated for this tract is theoretically 37,551 bp, and corresponds to the length between PSV V595 and the end of P6 arms.

About P7 and P8 palindromes, from our NGS data it was not possible to define the exact average length of the gene conversion tracts, which span from a minimum of 1 bp to a maximum of few kilobases.

The inter-chromosomal gene conversion

X-to-Y gene conversion in P8 palindrome

Due to the relevance of the inter-chromosomal gene conversion in modulating the genetic diversity of the MSY (Cruciani et al. 2010b; Trombetta et al. 2010, 2014), we evaluated the extent of the X-to-Y non-allelic gene conversion within the whole gametologous region of palindrome P8. To this aim, we performed four pair-wise alignments between human P8 arms and each of the 4 gametologous sequences on the X chromosome, each containing a copy of the VCX gene.

One of the consequences of a gene conversion event from the X to a single arm of palindrome P8 is to increase the number of differences between palindrome arms, thus we specifically searched for P8 SNPs with the derived allele corresponding to the gametologous base on the X chromosome. By this approach, we identified a minimum number of 22 X-to-Y gene conversion events across the phylogeny, involving 16 PSVs of the Y chromosome (Table 7; Additional file 3). Some positions were interested by a frequent inter-chromosomal conversion activity (V683, V687) and only for two cases we were able to define the exact donor from the X chromosome (V669, V687). These events increase the divergence between arms by creating new PSVs or re-introducing them after the Y-Y gene conversion has occurred, but at the same time makes X-Y copies more similar each other.

Region of P8	Length (bp) in GRCh37/hg19 (proximal-distal)	% arm-to-arm similarity	Sequenced bp	PSV	X-to-Y events per PSV	co-conversion
Additional flanks	2,846-2,246	91.3%	2,231-1,991	V669	1	-
VCY	741-741	100%	741-741	V120* V121*	1	yes
				V118*	1	-
				V671	1	-
				V672 V673	1	yes
				V119	1	-
X-Y gametologous region excluding flanks and VCY	12,476-12,476	100%	5,521-5,519	V675	2	-
				V676	2	-
				V109*	2	-
				V678 V679	1	yes
				V681	1	-
				V683	5	-
				V687	3	-

Table 7 (previous page): List of the X-to-Y gene conversion events for different portions of the X-Y gametologous region of palindrome P8. For each portion, the Y-Y similarity and the number of sequenced bases is reported.

Interestingly, by comparing the PSVs/sequenced-bp ratio among different regions of P8 palindrome, we found a significant difference between X-Y identity portion and the no-identity one, as described above, but in particular we found that the small VCY gene (741 bp) exhibits the highest relative content of variants (Figure 19), with 7 out of 8 PSVs interested by the X-to-Y gene conversion (V120*, V121*, V118*, V671, V672, V673, V119).

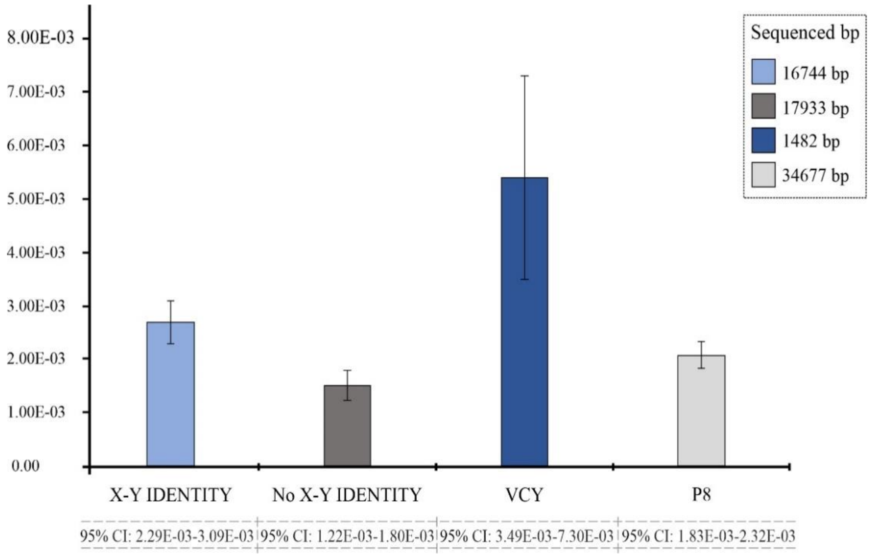

Figure 19: PSVs/sequenced bp ratio for different portions of palindrome P8. Below each column chart is reported the corresponding 95% CI. In the top right box are indicated the base pairs sequenced (in both arms) for different regions of P8.

Moreover, by comparing the X-Y identity region after excluding the VCY gene with the no-X-Y-identity portion of P8, we found a non-significant difference in the accumulation of PSVs

($p= 0.061$; Fisher Exact test). From our data, we concluded that the excess of PSVs observed in the VCY gene may possibly represent the cause of the heterogeneous distribution of PSVs between the gametologous region and the non-gametologous one, observed in P8 palindrome, confirming its behaviour as hotspot of inter-chromosomal gene conversion (Trombetta et al. 2010).

In addition, in agreement with the higher number of PSVs recorded together with the significantly higher mutation rate observed within the X-Y region of identity, we also observed a significantly higher human-chimpanzee sequence divergence within the X-Y identity region with respect to the rest of palindrome (3.36% vs 1.83%, $p < 0.0001$), suggesting that the X-Y gene conversion may play a role also in the independent evolutionary histories of these species.

In order to confirm these results and to investigate historical X-Y gene conversion events, we used the method described in Trombetta et al. (2014) to perform four-way sequence alignments of the highly-similar X-Y regions shared between sex chromosomes of human and chimpanzee reference sequences. In particular, we produced 4 different alignments of P8 proximal arm with each counterpart on the X chromosome. By this approach, we identified a total of 35 inter-species conversion sites (C-sites), 10 of which shared between 2 or more X chromosome sequences. Interestingly, for each alignment, we found that the C-sites were not equally distributed, recording the highest C-sites/kb content within VCX/VCY gene. This pattern makes the VCX/VCY gene a C-sites Enriched Region (CER, ≥ 4 C-sites/kb), confirming that conversion in this gene family has a deep evolutionary background and is probably involved in independent evolutionary histories of the orthologous genes.

X-Y gene conversion tract-length and rate estimate

By aligning the sequences of palindrome P8 with the four gametologous regions on the X chromosome, we estimated a minimum gene conversion tract ranging from 1 to 14 bp and a maximum possible length varying between 3 and 170 bp (Table 8). These X-to-Y gene conversion tracts estimated resulted to be shorter than the general estimates of Y-Y gene conversion tracts (Bosch et al. 2004; Hallast et al. 2013).

PSV in X-Y similar region of P8	X-to-Y events per PSV	Minimum tract-length (bp)	Maximum tract-length (bp)
V669	1	1	3
V120* V121*	1	11	44
V118*	1	1	170
V671	1	1	96
V672 V673	1	6	90
V119	1	1	107
V675	2	1	92
V676	2	1	83
V109*	2	1	38
V678 V679	1	14	134
V681	1	1	37
V683	5	1	21
V687	3	1	44

Table 8: Minimum and maximum length of X-to-Y gene conversion events identified in the gametologous portion of palindrome P8.

By using a slightly modified version of the method described in Cruciani et al. (2010b), we exploited the total minimum (41 bp) and maximum number (959 bp) of converted bases to estimate a gene conversion rate in the whole region of X-Y similarity of P8, sequenced in this study (16,744 bp). We obtained an X-to-Y gene conversion rate ranging between 8.32 × 10^{-10} (sd ± 0.43 × 10^{-10}) and 1.95 × 10^{-8} (sd ± 0.1 × 10^{-8}) events per base per year. This rate expresses the probability per year that a site is involved in an X-Y gene conversion event.

Similarly, we selectively calculated the X-to-Y gene conversion rate within the VCX/VCY gene, which resulted in a figure varying from 4.58 × 10^{-9} (sd ± 0.24 × 10^{-9}) to 1.16 × 10^{-7} (sd ± 0.06 × 10^{-7}) events per base per year. This value is significantly higher than the rate of the whole X-Y identity region analysed by us ($p < 0.0001$, Comparison of two rates test).

DISCUSSION

Through the use of a powerful tool, such as a robust phylogeny of the Y chromosome, together with high-resolution NGS analyses, we carried out the first unbiased study on the genetic diversity of human MSY palindromes, identifying many more PSVs than are present in the reference genome (GRCh37/hg19). This increased our ability to identify gene conversion events occurred during the recent human history, allowing the study of their molecular dynamics. In order to avoid possible interpretation issues due to sequence misalignment, we focused exclusively on the three "singleton" palindromes of the ampliconic MSY, i.e. P6, P7 and P8, characterized by a single repeat unit for each arm. The fact that palindromes analysed in the present study are also evolutionary conserved in chimpanzee, allowed a more accurate identification of PSVs and assessment of gene conversion features.

Gene conversion and structural variations

The most commonly used bioinformatics methods for the analysis of sequencing reads inadequately discriminate between genomic regions with almost complete sequence identity: the alignment programs do not incorporate information regarding the mapping of reads, therefore they cannot assign the right phase of paralogous variants identified with respect to a reference sequence. Moreover, this mapping distortion tends to affect downstream analyses, including detection of PSVs. In this regard, it has been necessary to carefully consider each case of reads mis-mapping (Figure 13) and to investigate them into a phylogenetic context, with the aim of identifying all the mutations and gene conversion events that shaped the diversity of MSY palindromes. Thus, we developed a new method for data analysis based on the assessment of the read distribution over the paralogous sites of palindromes, combined with the comparison of read depth between these highly similar sequences.

The sequencing depth analysis is of primary importance for the identification of ‘pseudo-hemizygous’ and ‘pseudo-homozygous’ deletions. In such cases, the read distribution over the paralogous sites alone would not be able to provide sufficient information about the exact genotype of a ‘pseudo-diploid’ variant, for example, if it is present as a ‘pseudo-homozygous’ genotype (i. e. T/T), in its ‘pseudo-hemizygous’ state (T/-) or ‘pseudo-homozygous’ (-/-) deleted form. Only through the depth analysis we will be able to distinguish among these 3 genotypes, for which we expect, respectively, a depth = 2N, depth = N and depth = 0, for each paralogous position.

Through this analysis, in our study we identified a single ~1.4 kb ‘pseudo-homozygous’ deletion within palindrome P6 of two A2-PN3 samples, which includes 3 PSVs found along the phylogeny (V540, V541, V542). We hypothesized that such deletion, firstly examined in silico, then experimentally validated by PCR and Sanger sequencing, has been generated along branch 9 of our phylogeny (Figure 12) by an intra-chromosomal homologous recombination (HR) between two 217 bp DRs on a P6 arm (Figure 20A-B-C). Then, the deletion has been transferred to the paralogous arm by a single gene conversion event (Figure 20D-E), causing the loss of a 1,393 bp sequence from both arms of palindrome P6.

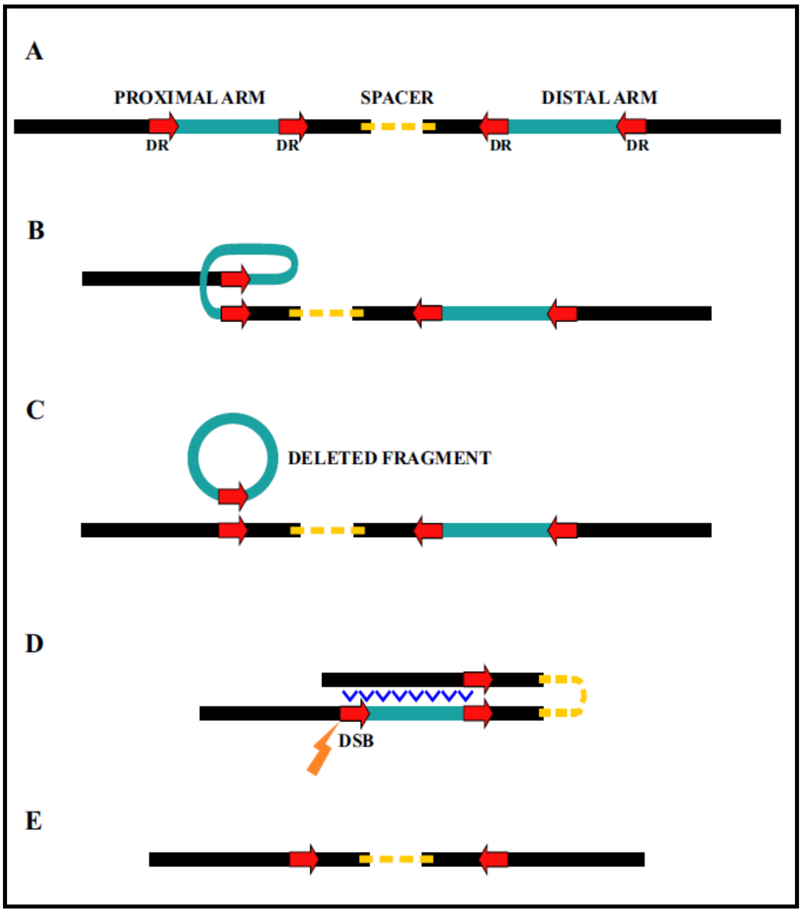

Figure 20: Proposed mechanism for the formation of the ‘pseudo-homozygous’ deletion in the A2-PN3 lineage. **A)** P6 palindrome structure before intra-chromosomal homologous recombination (HR), **B)** HR between direct repeats (DRs) on the proximal arm, **C)** The loss of the sequence between DRs (and of a DR itself), results in a ‘pseudo-hemizygous’ deletion, **D)** DSB formation and subsequent gene conversion from proximal to distal arm. **E)** Generation of the ‘pseudo-homozygous’ deletion after the arm-to-arm gene conversion.

It has already been supposed an involvement of gene conversion in the onset of CNVs on the human Y chromosome of cell lines (Shi et al. 2018) that agrees with our hypothesis, suggesting that the NAGC may cause deletions not exclusively in cell lines, more likely to be affected by the accumulation of mutations, but also in the germline. However, the other possible hypothesis explaining such observation points that the deletion arose on the paralogous arm of P6 by means of a double inter-chromatid crossing-over (CO), since both gene conversion and double CO produce indistinguishable structures (Chen et al. 2007). However, single inter-chromatid crossing-overs are rare since they could result in aberrant chromosomes with clinical consequences (Lange et al. 2009) and the probability that two occur in a few kilobases is quite low if we consider the steric footprint of the recombinative machinery (Shi et al. 2018). Thus, it is possible that the non-allelic gene conversion represents the eligible mechanism to justify such phenomenon.

According to this, it seems that gene conversion, in spite of maintaining the structural integrity of sequences by repairing DSBs, as previously hypothesized (van den Bosch et al. 2002; Rozen et al. 2003; Shrivastav et al. 2008), may be also involved in the “fixation” of deletions and the loss of genetic material from arms, suggesting the potential of Y-Y recombination as an evolutionary force capable of generating genetic erosion within ampliconic sequences.

Direction of the gene conversion

Absence of the ancestral bias and evidence for the GC-bias

By comparing orthologous (human-chimpanzee) sequences, Rozen et al. (2003) found an average inter-species divergence corresponding to 1.44% along palindrome arms, compared to the 2.2% of the spacer ($p < 0.05$).

The lower and statistically significant divergence between orthologous palindrome arms compared to that of spacers could be explained by two hypotheses: 1) gene conversion between arms may preferentially act by converting a new mutation to its invariant ancestral state, leading to the establishment of a low level of average divergence between orthologous sequences over time; 2) the mutation rate within arms is lower than the mutation rate within the spacer, that, as a consequence, will diverge from its ortholog more rapidly.

Rozen et al. (2003) assumes that the observed discrepancy between inter-species duplicated and unique regions is mainly due to a tendency for gene conversion to revert mutations to their ancestral state, thus preserving high similarity between paralogous and orthologous sequences. Accordingly, Hallast and colleagues (2013), through the analysis of 10 PSVs of the human P6 palindrome, found a weak significant excess of gene conversions towards the ancestral state, and the same was observed in gorilla and chimpanzee. In addition, more recently, by the analysis of 2.7 Mb of the ampliconic region of 62 subjects covering a little portion of the Y chromosome diversity, Skov and colleagues (2017) reported evidence for a gene conversion bias towards the ancestral state. However, from a more detailed analysis of their data here performed (data not shown, analysis based on S3 Dataset in Skov et al. 2017), such bias emerges from only 2 out of 8 MSY palindromes (P1 and P5), and it results to be completely absent in P6, P7 and P8.

From our analysis, we found no evidence for a gene conversion bias towards the ancestral state, suggesting that this non-allelic recombination mechanism may actually work without a specific direction, contradicting the hypothesis that Y-Y gene conversion evolved to retain the ancestral state of palindromic sequences. Indeed, we firstly recorded a non-significant difference between the number of ancestral and derived events in P6 and P8 palindromes, but a slightly significant excess towards the derived state in P7 palindrome was observed. It could be thought that this result is mainly influenced by the underestimate of the to-ancestral gene conversions that cannot be detected in case they occur on the phylogenetic branch where the variant arose, masking the mutational event. Therefore, we eliminated as many derived events that would not have been detectable if they had occurred towards the ancestral, but also after such calibration, we found no significant differences between the number of to-ancestral and to-derived gene conversions in all the 3 palindromes here analysed, both separately and combined (p-values > 0.05).

Thus, it seems that the ectopic gene conversion, at least in palindromes P6, P7 and P8, does not work with a specific direction. On the contrary, a preferential direction of Y-Y recombination emerged from the analysis of a bias towards the fixation of specific nucleotides. We found a significant excess of conversions fixing GC bases over AT in P6 and P8 palindromes, that is in line with the GC-conversion bias already abundantly observed in several taxa including mammals (Galtier et al. 2001; Galtier 2003; Kudla et al. 2004; Duret and Galtier 2009; Hallast et al. 2013; Lartillot 2013; Skov et al. 2017).

In order to assess the reasons for the conflict between our results and Hallast et al. (2013) results about the ancestral/derived direction of gene conversion, we point that, differently from us, Hallast et al. (2013) analyse PSVs that fall mostly into repeated elements discarded by us, which could recombine via NAHR with highly similar repeats present on both sex-chromosomes and autosomes (Trombetta et al. 2016), thus affecting Y-Y gene conversion and further complicating the interpretation of the diversity pattern of Y-linked palindromes. In addition, after excluding PSVs that are informative for the GC-bias (GC-bias correction), we found no difference between the number of ancestral and derived events, neither under the assumption that all events are independent ($p = 0.9028$, Chi-square test) nor after removing the putative co-conversions ($p = 0.7728$, Chi-square test). Moreover, all the remaining PSVs (those informative for the GC-bias) are converted towards the ancestral or derived state according to the conversion towards the G or C nucleotide, and most of them, returning to the ancestral state, have as ancestral base a G or C nucleotide.

Accordingly, from our study it seems that the GC-bias is the main driving force of the gene conversion event. Indeed, in addition to the higher number of GC-conversions (both ancestral and derived) with respect to the conversions fixing AT nucleotides (see Results), we found that also among GC-biased events, there is no difference between the ancestral and derived ones, confirming the absence of a specific directional bias and making the GC-bias the only bias apparently existing.

The GC-biased gene conversion is probably due the higher incidence of DSBs at the weaker A-T base pairs, with respect to paralogous G-C base pairs, because of their lower energy due to the presence of only two hydrogen bonds. Indeed, a conversion bias is expected when one haplotype is more prone to the formation of DSBs (Duret and Galtier 2009), thus, when we are in presence of A/G or T/C PSVs, as a consequence of the double-strand break in proximity of the A or T nucleotides, the gene conversion will necessarily draw the information from the paralogous G or C positions, which will act as donor sequences of the conversion event.

Dynamics of palindromic sequences: new insights on gene conversion and mutation rate

Accurate estimate of the gene conversion rate

The estimate of gene conversion rate with the method reported for the first time in this study represents an advancement with respect to previous studies (Hallast et al. 2013; Skov et al. 2017). In particular, our focus is on accurate time estimate (the denominator of the rate), which is an item of primary importance in a rate calculation. In previous studies the gene conversion rate has been calculated, roughly, as the ratio between the number of gene conversion events and the whole time spanned by the phylogeny. However, this would mean estimating the gene conversion rate also in a time when the conversion events cannot be detected, such as when they occur on branches of phylogeny not affected by a mutational event, whereas the observation of a gene conversion presumes the existence of a PSV (‘pseudo-heterozygous’ state).

Indeed, if we consider a simplified phylogeny joining 5 chromosomes, A, B, C, D, E, and suppose that two mutations occur on A and D branch, respectively, generating two PSVs that will be subsequently converted to the derived state by two independent gene conversion events (Figure 21-1), by calculating gene conversion rate with the method reported in literature, we would obtain, for this specific PSV, a simplified rate equal to *2/whole time of phylogeny* (up to the AE node).

Similarly, in a different scenario, a single mutational event generating the PSV occurred along the root of the tree, making all the 5 chromosomes being characterised by a ‘pseudo-heterozygous’ state. Then, we observe two to-derived gene conversion events, respectively on branches A and D (Figure 21-2). According to the method described in literature, we would obtain the same calculation of the gene conversion rate, equal to *2/whole time of phylogeny*, because this method is not able to discriminate between the two possible scenarios and does not take into account the time of existence of each PSV.

On the contrary, it is reasonable to think that, in the two cases just described, time estimate provides different contribution to the gene conversion rate calculation, since, in case 1, the two PSVs created by different mutational events have a shorter time (equivalent to the times of branches A and D) to be converted with respect to case 2, where gene conversion events may occur (and be observed) in a wider time frame, corresponding to the whole time of phylogeny. Thus, from the observation of an equal number of conversion events within two PSVs existing since different times, we can deduce different rates of conversion depending not only on the number of events observed, but also on time. In this regard, we introduced the estimate of an "observed gene conversion rate”, which takes into account the actual time when the gene conversion events can be detected; thus considering the time with which each PSV may really contribute to the observed gene conversion rate.

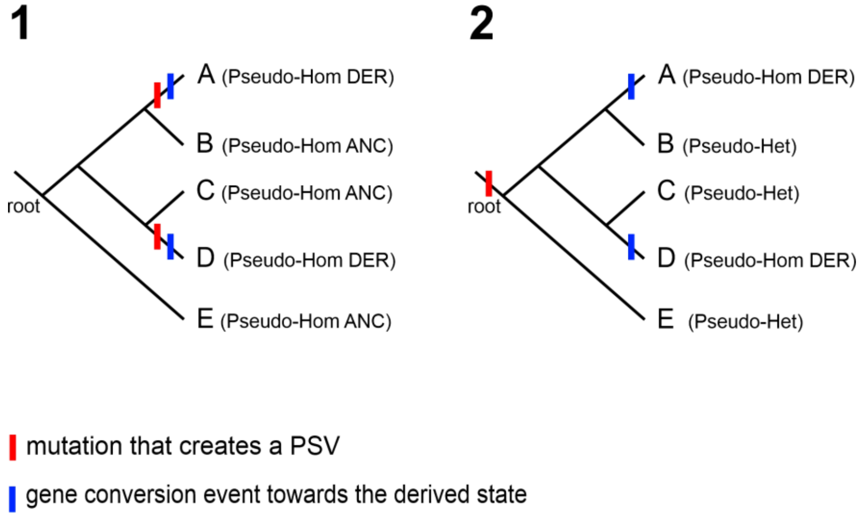

Figure 21: Simplified phylogenetic tree joining 5 chromosomes. **1)** A single mutational event (red rectangle) and subsequent to-derived gene conversion (blue rectangle) occurring on branches A and D, respectively. **2)** Single mutational event occurring at the root of the tree, generating a PSV (‘Pseudo-Heterozygous’ state) in all 5 chromosomes. Then, two distinct gene conversion events occurred on branch A and D, respectively, leading to a ‘Pseudo-Homozygous’ derived genotype in A and D chromosomes.

Our results about the palindrome-specific gene conversion rate suggest that P6 exhibits the highest rate observed among the three palindromes analysed, showing a stronger tendency to dilute mutations by gene conversion, whereas the evolutionary dynamics of P8 palindrome, which exhibits the lowest gene conversion rate (see Results), seem to be strongly affected by gene conversion from the X chromosome, as suggested by the highly different rates observed between the gametologous and non-gametologous region.

Evaluation of mutation – gene conversion equilibrium

Since the mutational event creates new PSVs between the arms of the palindrome, whereas gene conversion erases those differences, it is possible that a steady-state equilibrium in the diversity between arms is established, where the mutation rate (that increases the diversity between arms) is counterbalanced by the gene conversion rate (that dissolves this diversity). Assuming the existence of a balance, it is possible to calculate the conversion rate per base per year (Rozen et al. 2003) using an averaged-π (specifically calculated for each palindrome) and the palindrome-specific mutation rate (see Results). We obtained a gene conversion rate equal to 7.52×10^{-6} (sd ± 0.39×10^{-6}), 5.74×10^{-6} (sd ± 0.3×10^{-6}) and 5.86×10^{-6} (sd ± 0.3×10^{-6}) events per duplicated nucleotide per year, respectively for P6, P7 and P8 palindromes.

When these gene conversion estimates, based on the equilibrium assumption, were compared with the actual gene conversion rates, estimated with our method, we found that:

1. Only in case of P6 palindrome, the gene conversion rate estimated under the hypothesis of a steady-state balance is in the same range of variability of the gene conversion rate that we calculated without considering this hypothetical equilibrium (Table 9). This may suggest the establishment of a mutation/conversion balance that maintains an average level of diversity between P6 arms over time, making this palindrome the most stable among the three here analysed;

2. For P7 and P8 palindrome we found no correspondence between the two gene conversion rates estimated with different methods (Table 9). In particular, the observed gene conversion rate of P8 resulted to be about 2.5 times lower than that estimated under the hypothesis of balance, suggesting that the evolutionary dynamics of P8 palindrome are governed not only by Y-Y recombination, but are strongly influenced by the inter-chromosomal gene conversion.

Pal	**Observed gene conversion rate (per base per year)**	**Expected gene conversion rate under the hypothesis of mutation-gene conversion equilibrium**
P6	4.42-9.38 × 10^{-6}	7.14-7.92 × 10^{-6}
P7	3.4-4.69 × 10^{-6}	5.44-6.04 × 10^{-6}
P8	1.98-2.2 × 10^{-6}	5.56-6.17 × 10^{-6}

Table 9: Comparison between the observed gene conversion rate (this study) and the expected gene conversion rate under the hypothesis of mutation-conversion balance (right column).

Implication of the mutation rate in the evolution of palindromes

Our estimates for the mutation rate of P6, P7 and the non-gametologous X-Y region of P8 seem to be consistent each other, although they are not in the same range of variability of mutation rate estimated by Helgason et al. (2015). On the contrary, the mutation rate of P8 (as a whole, considering both the gametologous and non-gametologous portions) resulted to be higher than that estimated for the other two palindromes and consistent with that calculated by Helgason et al. (2015). This could be related to the fact that Helgason et al. (2015) analysed MSY palindromes as a whole, obtaining an average mutation rate for the entire palindromic region, whereas different palindromes could actually exhibit different mutational dynamics.

About spacer-specific mutation rates, we found that all of them were coherent with the rate of the haploid X-degenerate region estimated by us ($p > 0.05$). In particular, the mutation rate obtained within the spacer of P6, the palindrome for which we have more sequencing data, resulted to be significantly higher than both mutation rates estimated within arms (see Results). The same trend is observed for P7 and P8 palindromes, although the difference between spacer and arms in this case does not reach the threshold for statistical significance, possibly due to the inconsistent number of bases sequenced within the spacers (corresponding to non-repetitive elements, 1,680 and 1,886 bp for P7 and P8, respectively), and the low amount of mutations identified.

Interestingly, the discrepancy between ‘pseudo-diploid’ and haploid regions observed in P6 palindrome may explain the lower divergence between the orthologous (human-chimpanzee) P6 arms compared with the divergence of spacers, previously observed (Rozen et al. 2003), rather than hypothesising that gene conversion is acting to revert new arising mutations to their ancestral state (Rozen et al. 2003; Hallast et al. 2013). However, if an enhanced mutation rate within the spacer exists, it could be related to the gene conversion mechanism itself: during the recombination and pairing of paralogous arms, palindromic sequences form cruciforms as intermediate structures, characterised by a four-way DNA junction and stems containing single-stranded loops (Pearson et al. 1996). In this condition the spacer forming the loop would remain for a certain time in a single-stranded state, where it could be more likely exposed to mutational phenomena.

More in general, despite the absence of protein-coding genes in P6 and P7 palindromes, evidence of functional elements overlapping P6 and P7 sequences has been recently reported (Chechova et al. 2020), so a role of natural selection should be also considered. In this context, natural selection can cause the directional removal of variants appearing during the evolutionary time scale separating man and chimpanzee, actively participating in shaping the diversity of inter-specific palindromes. When a new variant arises in one arm, even if slightly harmful, it could escape the action of natural selection since its hypothetical function is governed by the unmodified base on the other arm of the palindrome. In this context, gene conversion may act towards the retention of the ancestral base, restoring the invariant state and making it invisible to the selection or, on the contrary, may act by fixing the derived state, leading to the establishment of a harmful variant on both arms of the palindrome, so that purifying selection will probably act by eliminating such variant.

Thus, even though a significant difference between gene conversion towards the ancestral and derived state is not observed in the three palindromes here analysed, the effects of Y-Y recombination over a long evolutionary time are those of the ancestral, which could favour the maintenance of inter-species sequence identity also in absence of a direction bias.

In conclusion, the higher inter-species similarity between orthologous palindrome arms with respect to the spacer may be consequence of:

1) A higher mutation rate of spacers compared to palindrome arms;

2) Purifying selection that eliminates ‘pseudo-homozygous’ derived and potentially deleterious variants, favouring the ancestral ones over time.

Role of the X-to-Y gene conversion in the evolution of P8 and VCX/VCY genes

In our study we found that the Y-Y non-allelic homologous recombination does not seem to exhibit equal trends among different palindromes, suggesting that palindromic sequences may be involved in independent evolutionary histories. Indeed, both gene conversion and mutation rates obtained from our analyses bring out dissimilar behaviours among P6, P7 and P8. In particular P8, the only palindrome hosting a coding gene, exhibits evolutionary dynamics that differ from those of P6 and P7. However, with a deeper level of investigation, we found that also different regions within P8 arms show highly different evolutionary paths. The existence of NAHR hotspots within paralogous regions has already been suggested since long time, as it has been observed that recombination in such sequences does not occur uniformly, but is concentrated in 'hotspots' characterised by high recombination rates (Lupski 2004). In particular, local ‘hotspots’ are defined as short genomic regions where strand exchanges are more common than elsewhere, despite different NAHR hotspots do not seem to share common sequence motifs (Lupski 2004).

For a long time, it was believed that recombination between human sex chromosomes was limited to pseudo-autosomal regions. However, it has been recently shown that non-allelic recombination is also active in paralogs of the MSY in the form of X-to-Y gene conversion and, in particular, within P8 palindrome it has been demonstrated the existence of a proper gene conversion hotspot involving the VCX/VCY gene (Trombetta et al. 2010).

By investigating the X-to-Y gene conversion role in the whole X-Y region of P8, sequenced by us, we found 22 gene conversion events from the X chromosome, that contribute to increase the Y-Y divergence through the creation of new PSVs. In particular, the higher number of PSVs recorded within the gemetologous region has to be referred to the hyper-accumulation of PSVs within the VCY gene ($p < 0.05$), mostly introduced by the X chromosome.

This suggests that the sequence landscape of palindrome P8 can be modulated by the transfer of genetic information from the X chromosome, that affects in particular VCX/VCY genes. This may lead to the hypothesis that this gene family can follow autonomous evolutionary trajectories with respect to the rest of palindrome.

To date, the biological function of proteins encoded by VCX/VCY genes is still unknown. Some studies suggest they may have a role in pathogenic phenomena, since they were found to be mutated in some kind of tumours (Taguchi et al. 2014; Deng et al. 2018), or they were thought to play a role in the assembly of ribosomes during spermatogenesis (Zou et al. 2003). However, these biochemical properties do not clarify the biological function of VCX/VCY members in the whole organism and no deletions that remove both VCY gene copies have been still reported in humans (Shi et al. 2019), thus genetic analyses have not helped to define the gene function yet. Other studies suggest an origin of the VCY gene after the divergence from macaque, about 25 million years ago (Hughes et al. 2012); however, more recently, it has been proposed the appearance of VCY copies in human-chimpanzee-bonobo common ancestor, with the subsequent lost in bonobo lineage (Chechova et al. 2020).

Thus, it seems that only human and chimpanzee species retain copies of VCX/VCY gene family, that we found to be involved in historical X-Y gene conversion, identified by four 4-way sequence alignments of human-chimpanzee orthologs. The excess of intra-species conversions (C-sites) recorded within VCX/VCY gene in this study (8 out of 35 C-sites found in less than 1 kb) suggests that, in at least one of the two species, mutation and subsequent gene conversion occurred towards the fixation of the derived state of variants, necessarily leading to the differentiation of the orthologous gene copies, which possibly drives the diversification of the orthologous proteins, maybe in order to establish some species-specific function. This was also in line with the significantly higher inter-species divergence that we observe between orthologous VCY copies (5.03%), compared with the rest of palindrome (Table 10).

P8 TOT	P8 X-Y identity	P8 no-X-Y-identity	VCY	Spacer
2.46%	3.36%	1.83%	5.03%	2.84%

Table 10: Human-chimpanzee sequence divergence for different regions of P8 palindrome and P8 as a whole.

In agreement with this observation, it has been previously demonstrated an extraordinarily high nucleotide sequence divergence between human and chimpanzee coding regions of VCY, compared with the introns (Hughes et al. 2010), that corresponds to a dN/dS ratio > 1. Although it was indicative of a positive selection for rapid amino acid-sequence divergence, this dN/dS ratio mainly resulted from structural alterations at the 3′ end of the coding region, suggesting that this gene may be not-essential in chimpanzee (Kuroki et al. 2006). Indeed, the human-chimpanzee VCY orthologs may begin to diverge either by the accumulation of non-synonymous mutations in humans, that, as a result of positive selection, may confer an advantageous role to the human species, or by the accumulation of mutations in the chimpanzee VCY, that leads to the loss of its function. However, the maintenance of an active ORF and gene expression in human (Lahn and Page 2000; Lahn et al. 2001) suggests functional importance in our species, which needs to be clarified.

From our analysis we found that the X-to-Y gene conversion rate in the whole gametologous region resulted to be about 18-20-fold lower than previous estimates reported for the VCX/VCY gene family (Cruciani et al. 2010b). This is possibly due to the clearly higher X-to-Y exchange rate in the small VCY genes. Indeed, by selectively calculating the X-to-Y gene conversion rate in VCY, we obtained a significantly higher figure with respect to the conversion rate of the X-Y region as a whole, and resulted to be in the same range of variability of the estimates reported by Cruciani et al. (2010b) for the same region. Moreover, the accelerated dynamics of X-to-Y gene conversion found in the VCYs seem to suggest the necessity of preserving high similarity between X-Y gene copies in humans in form of concerted evolution, which is coherent with previous hypotheses that members of the VCX/VCY protein family can work together by complementing each other in functions involved in spermatogenesis (Lahn and Page 2000; Van Esch et al. 2005).

In light of this, the abundant X-to-Y gene conversion that we found between VCX and VCY genes may provide a valid contribution against the decay of this important gene family in human species; on the contrary, it may be indicative of loss of function in chimpanzee.

It is possible that the X-to-Y gene conversion is a biological phenomenon distinct from the Y-Y recombination in palindromes, since it affects sequences characterised by an average lower similarity, which has been estimated in this study of 93.3% (BLAT results). Indeed, the minimum homology between the interacting sequences has been demonstrated to be always >92% and usually >95% (Chen et al. 2007). Although there is evidence for consistent inter-chromosomal exchanges, it could be not necessarily due to the gene conversion itself, but possibly to the presence of fragile sites where double-strand breaks are more likely to occur and where gene conversion, given its higher incidence rate compared to crossing over (Chen et al. 2007), is more likely to resolve them.

On the contrary, Y-Y gene conversion seems to be a constitutively active mechanism in Y chromosome palindromes, driven by the particularly high sequence identity between arms (varying from 99.94% to 99.997%). This is also in line with the generally longer Y-Y gene conversion tracts (Bosch et al. 2004; Hallast et al. 2013) compared to the average shorter tracts involved in X-Y gene conversion (Rosser et al. 2009; Trombetta et al. 2010, 2014), which resulted to be equal to an average of 38.5 bp in the present study. Indeed, the minimal processing segment for an efficient gene conversion has been estimated to be in the range of 337–456 bp in humans (Reiter et al. 1998).

Hypotheses on the biological role of palindromes and Y-Y gene conversion

Both P6 and P7 palindromes do not harbour coding genes, so it is reasonable questioning what are the reasons that led to their preservation over time. In particular, P6 palindrome seems to be the palindrome exhibiting the highest percentage of sequences shared with chimpanzee (where it corresponds to C19 palindrome) and both P6 and P7 sequences have been found in multi-copy state in most great ape species (Chechova et al. 2020). Moreover, according to our data, P6 seems to be the most evolutionary stable palindrome of the MSY in humans, since it is the unique amplicon analysed by us interested by a steady-state mutation-conversion equilibrium and that maintains the ancestral state of two branches in 1,216 samples covering a high diversity of the Y chromosome tree (Teitz et al. 2018). Therefore, it is plausible that P6 and P7 conservation is driven not by spermatogenesis-related genes, but by other elements regulating gene expression. Indeed, open-chromatin markers and protein-binding sites that could act as regulatory elements of gene expression have been recently identified in both these palindromes (Chechova et al. 2020).

More in general, considered the ubiquity of ampliconic sequences on mammalian Y chromosomes, it has been proposed that the amplification mechanism itself confers functional benefits, such as the rescue of deleterious mutations through gene conversion and the dosage of ampliconic genes. Nevertheless, another hypothesis already partially addressed is that palindromes may allow ampliconic genes to escape the well-known process of meiotic sex chromosome inactivation (MSCI) (Teitz et al. 2018).

We hypothesize that palindromic genes, that are necessary for the maturation process of spermatozoa, need to be expressed before the completion of meiosis. Thus, by pairing with themselves during pachytene, they may recombine intrachromosomally without structural problems, in a way that is enhanced by their organization as inverted repeats itself. In this view, it is possible that gene conversion may not be a process evolved "per se", able to maintain exclusively the structural integrity of gene sequences, but it could also be (or exclusively be) the consequence of the more extensive process of meiotic recombination, where pairing and recombination between ampliconic genes could ensure their expression by escaping the inactivation of the entire Y chromosome.

CONCLUSIONS

It is well known that palindromic sequences are a common feature of the sex-limited chromosome of different species and it has been hypothesized that non-allelic gene conversion within palindromes evolved to preserve the MSY ampliconic gene content from degeneration, by withstanding new deleterious mutations through the return to their ancestral state (Rozen et al. 2003; Hallast et al. 2013; Skov et al. 2017). However, this evidence is based on poor population data, and despite the general importance of non-allelic gene conversion emerged from previous analysis on the MSY, from our study we can conclude that:

- If on one hand gene conversion has a role in maintaining the structural identity of palindromic sequences, on the other hand it seems to play a role in the fixation of deletions and the irreversible loss of genetic material from palindrome arms;

- Gene conversion activity is not biased towards the retention of the ancestral state in all MSY palindromes. The to-ancestral bias does not emerge neither from gene-free palindromes (P6 and P7), nor from P8, which hosts a protein-coding gene. Rather, gene conversion direction seems to be governed by the GC nucleotide fixation-bias, which could be related to energetic features of the chemical process itself;

- We found higher mutation rates in the spacers compared to palindrome arms. This difference in mutation rate may represent the true cause of the previously observed higher human-chimpanzee spacer divergence with respect to the arms, without the need to invoke a Y-Y recombination bias towards the ancestral state.

- P8 palindrome is characterized by different functional regions that follow independent evolutionary dynamics. In particular, the X-Y gametologous portion, involved in both intra- and inter- chromosomal gene conversion, shows high mutation rate and low gene conversion frequency. On the contrary, the non-gametologous portion of P8 exhibits the highest gene conversion rate observed among palindromes here analysed and a mutation rate that is in line with that of P6 and P7.

- We found evidence for X-to-Y gene conversion within the whole gametologous region of P8 palindrome and we confirmed the presence of a hotspot of X-to-Y gene conversion in the VCY gene, which suggests the necessity to diversify orthologous gene copies, still preserving high similarity between gametologous genes, for some species-specific function.

MATERIALS AND METHODS

The sample

Selection of the sample set

The 157 Y chromosomes to be sequenced for palindromic regions (P6, P7 and P8) were selected from two different datasets on the basis of their genetic affiliation, determined in previous studies (D'Atanasio et al. 2018; Finocchio et al. 2018). They were selected from a collection of more than 5,000 samples of our laboratory in order to: i) maximize the genetic diversity within the phylogenetic tree; ii) cover the greatest possible time range.

Sample quality and quantity control

The DNA to be sequenced for palindromic regions was obtained from saliva, peripheral blood and cultured cells.

Target sequencing required specific quantity and quality parameters: 1) absence or low amount of DNA degradation; 2) quantity ≥ 3 μg; 3) concentration ≥ 37.5 ng/μl; 4) purity: A260/280 = 1.8-2.0. The quantification and purity were assessed using a Nano-Drop 1000 spectrophotometer and Qubit 4 Fluorometer, both produced by Thermo Fisher Scientific. Degradation was evaluated by means of an electrophoretic run on a 1% agarose gel. Overall, 98 samples passed all necessary criteria; 59 samples with low amount of DNA underwent a whole genome amplification (WGA) using the GenomiPhi V2 DNA Amplification kit (GE Healthcare). The WGA uses random short primers for the amplification of the whole genome, starting from nanogram quantities of DNA and resulting in microgram quantities of amplified products. The main disadvantage of this technique is the possible non-homogeneous amplification of different genomic regions.

Phylogenetic analysis of Y chromosomes

SNP calling and filtering of variants in the X-degenerate region

In order to reconstruct a stable and reliable phylogeny of our samples, we used the publicly available 157 sequences of the X-degenerate unique region (D’Atanasio et al. 2018; Finocchio et al. 2018). We also included in the analysis 4 additional ancient specimens precisely radiocarbon-dated (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015) for an accurate time estimate. By exploiting the .bam files of all the 161 subjects, we performed a new SNP calling using the same 3,328,701 bp of the X-degenerate region sequenced in D’Atanasio et al. (2018).

The filtering criteria adopted for the X-degenerate region were the same of D’Atanasio et al. (2018). The VCF parameters considered were the quality (“QUAL” field), the depth (“DP” field) and the number of reads with the reference or the alternative base (“DP4” values). Using the information contained in the DP4, we calculated the FilDP4 parameter, used for a more accurate filtering:

$$FilDP4 = \frac{Number\ of\ reads\ with\ ALT\ base - Number\ of\ reads\ with\ REF\ base}{Total\ number\ of\ reads}$$

We directly discarded variant positions with FilDP4 ≤ 0.3 and retained all the SNPs with FilDP4 > 0.8 and QUAL ≥ 100. In the other cases, we considered the phylogenetic context: for SNPs shared among samples belonging to the same haplogroup, we applied less severe criteria (FilDP4 ≥ 0.6 and the number of ALT reads ≥ 2), while we discarded the private positions with DP < 2 or FilDP4 and DP less than 0.4 and 4, respectively. The remaining cases were manually checked in the alignment files. We excluded mutations occurring at closely spaced positions (less than 20 bp) that may be indicative of structural rearrangements.

Phylogenetic tree reconstruction

The maximum parsimony tree of the 161 samples (157 modern DNAs and 4 ancient DNAs) was reconstructed by means of MEGA (Tamura et al. 2011; Kumar et al. 2016) after generating a .meg input file containing the 7,240 biallelic polymorphism identified in the whole sample set. Since the mutational events of branch 1 could not be assigned univocally to A00 or A0-T haplogroups, the root of the tree was positioned at midpoint by default. We used the Network software to produce a median joining network, submitting a .rdf file as input, and to calculate the specific rho values of nodes, which depend on the mutation rate of the phylogeny. The .out file returned by the program was manipulated to obtain the list of mutations for each branch and the positions of recurrent variants. These latter, together with the tri-allelic ones, were also checked in the .bam files of the samples of interest.

Mutation rate estimate within the X-degenerate region

The mutation rate for the ~3.3 Mb analysed was estimated by means of BEAST software (Drummond and Rambaut 2007). The input was a NEXUS (.nex) file, containing the list of the variable positions for all the 161 subjects and the structure of the maximum parsimony tree in the newick format. At first, the input was loaded onto BEAUTY suite, assigning to the four ancient specimens the calibrated radiocarbon dates in years before present (YBP) (Table 11). We used a GTR (general time reversible) nucleotide substitution model under a strict clock, an expansion growth model for the population size, and other flat priors as described in Trombetta et al. (2015a).

The output was checked with the Tree Annotator and Tracer platforms.

Sample	Radiocarbon Date (YBP)	Reference
Loschbour	8,055	Lazaridis et al. 2014
Bichon	13,665	Jones et al. 2015
Kotias	9,712	Jones et al. 2015
Ust'-Ishim	44,890	Fu et al. 2015

Table 11: C^{14} time estimates of four publicly available ancient Y chromosomes included in the phylogeny.

The mutation rate obtained and its standard deviation (sd) have been then calibrated for the 7,240 variants found in 3,328,701 bp analysed in the unique region, in order to estimate the specific mutation rate of our phylogeny, corresponding to a mean of 1 mutation every ~406.6 years.

Dating

Since we knew the mutation rate of our phylogeny, the dating of nodes of the tree and the related standard deviations were calculated using the rho statistics (Forster et al. 1996; Saillard et al. 2000).

The rho (ρ) value is an estimate of the average number of different sites between a set of sequences and their common ancestor, and it is measured as the average number of mutations downstream the node to be dated. This parameter is linearly related to mutation rate and time, according to the following equation: $\rho = \mu \times t$ (Jobling et al. 2004), assuming the constancy of the rate across the tree branches. The rho values were assessed by means of Network software (Bandelt et al. 1999) and used to estimate the time of most phylogeny nodes. The ages of the root and of some deep nodes of the tree (1, 5, 7 and 12) were calculated manually.

Target Next Generation Sequencing of palindromes

Selection of the targeted palindromic regions

To selectively target P6, P7 and P8 palindromes for the Next Generation Sequencing, we referred to the Y coordinates of the UCSC Genome Browser, Human Feb. 2009 - GRCh37/hg19 Assembly (Table 12).

At first, we selected ~0.4 Mb of the MSY ampliconic region, also including the spacer of the three palindromes. The final number of bases to be sequenced decreased to about 137 kb/sample after discarding the interspersed repeated elements. In order to obtain exclusively the coordinates of the non-repetitive palindromic regions, we used the "Table browser" tool of the UCSC Genome Browser.

Palindrome	Start position (GRCh37/hg19)	End position (GRCh37/hg19)	Span (bp)
P6	18271432	18537677	266,246
P7	17986738	18016824	30,087
P8	16093532	16172355	78,824

Table 12: Human Y chromosome coordinates of P6, P7 and P8 palindromes selected for the target sequencing.

Targeting and library preparation

The 157 DNA samples were analysed by the BGI-Tech of Hong Kong, which performed the targeting, library preparation, sequencing and alignment steps of the next generation sequencing.

The targeting consists in the enrichment of the reaction broth with the selected regions of the Y chromosome. In this way, the

DNA fragments representing the genomic regions under study will be present in high concentration within the library and will be specifically sequenced, avoiding the non-specific analysis of the whole genome. This procedure has been carried out using the NimbleGen pipeline, produced by Roche.

More precisely, genomic DNA was sheared by means of a Covaris ultrasonicator in order to obtain 200-300 bp DNA fragments. Adapters are subsequently attached to the ends of these fragments and the DNA is deposited onto a NimbleGen chip, on which probes of about 200 bp specifically recognize the target regions of DNA. In this way, the DNA of the region of interest hybridizes with the probes on the chip, while the remaining DNA will be removed. The probes excluded almost completely the repetitive elements, capturing a total of ~137 kb corresponding to the three palindromes under study.

For the library preparation, the enriched DNA is immobilized on a solid support, the flow cell, where there are 8 longitudinal channels, the lanes. Within each lane, small oligonucleotides are covalently bounded to the adapters attached to the end of the fragments to be sequenced. The introduction of DNA inside each lane allows the hybridization with the oligonucleotides fixed on the flow cell, after which, a clonal amplification reaction is performed to produce clusters of molecules to be sequenced.

Sequencing and alignment

The sequencing step has been performed by means of an Illumina Hi-Seq 2500 platform, producing a ≥ 50× mean depth for the targeted palindromic sequences.

The raw output was refined discarding low-quality reads and contaminations with adapters. The sequences of each subject were aligned to the human reference sequence (Human Feb. 2009 -

GRCh37/hg19 Assembly) by means of the BWA (Burrows-Wheeler Aligner) software (Li and Durbin 2009), producing an alignment file .bam format (Li et al. 2009), corresponding to the binary version of the .sam (Sequence Alignment Map) file (Li et al. 2009). The .bam file have been visualized by means of IGVtools (Integrative Genome Viewer tools).

Since our NGS analysis concerns duplicated regions of the genome, we are aware that each read will be mapped against both arms of the palindrome in the reference genome, thus the alignment step will require particular caution (see Results). For this reason, we could not discard low mapping-quality (MQ) reads, thus we also considered reads with MQ = 0.

Analysis of the sequencing data

Sequencing depth analysis

We performed 4 boundary-specific PCRs for each sample to test the presence of both proximal and distal arms of P6, P7 and P8 in the whole sample set. The primer pairs were designed with Primer3 software (v. 0.4.0) and selected in order to overlap the sequences between palindrome arms and unique regions.

To obtain reliable data for our 157 samples we performed the bioinformatic analysis of depth (DP) in both spacer and palindrome arms. At first, we extracted raw DP values from all positions of captured palindromes from each .bam file, using the SAMtools platform (Li et al. 2009; Li 2011), after taking an additional step for the removal of repetitive elements (using the “Table browser” tool of the UCSC Genome browser), not completely removed during the targeting phase. Due to the problems related to the read mapping, some positions may exhibit DP = 0, thus we applied a specific command line of SAMtools platform to include this information. For each sample we also calculated the average DP value for the ~3.3 Mb of non-repetitive regions of the MSY, whose .bam files were publicly available (D'Atanasio et al. 2018; Finocchio et al. 2018), and we used this value to standardize the per base DP values obtained within palindromic regions.

From the standardized DP values, we extracted the Exponential Moving Average (EMA) along the entire captured region, using 100-bp sliding windows 1-bp moving, by means of the "TTR" package provided by R tool.

Detection of putative duplications/deletions

In order to investigate possible structural variants within the palindromes under study, we specifically selected blocks of sequences with average EMA values <1.5 and >2.5 to detect deletions and duplications, respectively. Indeed, in absence of recombinative events that change their structural identity, palindromic regions should be characterized by standardised EMA values about 2 times higher than EMA of unique regions, such as the spacer (expected to be ~1).

Based on our outcomes, in the subsequent steps we decided to focus on the more reliable 'pseudo-homozygous' deletions resulted from the depth analysis. The identification of continuous clusters of positions with average EMA ≤ 0.1 in at least one subject were marked as putative 'pseudo-homozygous' deletions. We also inspected blocks of sequences with average EMA values <1.5 and >2.5 that were present in 2 or more phylogenetically related samples, since they may be indicative of a single recombinative event. All the selected blocks were subsequently checked in .bam files of samples of interest and validated by Sanger sequencing.

Variant calling within palindromic regions

The variant positions within palindromic sequences (arms and spacer) were identified aligning the sequences of the 157 samples to the Y chromosome reference sequence (Human Feb. 2009 - GRCh37/hg19 Assembly), using the SAMtools platform (Li et al. 2009; Li 2011). The output was a VCF (Variant Call Format) file for each sample, from which we filtered out the indels.

Filtering of variants within palindrome arms

In order to discard false positive variants, we applied some filtering criteria based on the 'pseudo-diploid' features of palindromic regions. The parameters took into account some information contained in the .vcf files, such as the "DP field" of each variant and the information about the number of reads with the alternative (DP_{ALT}) and reference base (DP_{REF}), both included in the "DP4" values within the "FORMAT field". Firstly, we applied the following filtering:

- If $DP \geq 2$ and $DP_{ALT} \leq 2$: the variant is discarded;
- If $DP_{ALT}/DP_{REF} < 0.1$: the variant is discarded.

After these steps, we have further refined the list of variants by calculating the Fl parameter, used for a more accurate filtering:

$$Fl = \frac{Number\ of\ reads\ with\ the\ ALT\ base}{Total\ number\ of\ reads\ (ALT + REF)}$$

We directly eliminated variant positions with Fl value < 0.1, since they probably represent false positive calls and we directly retained all positions with Fl value ≥ 0.9, assigning them as 'pseudo-homozygous' variants. About positions showing Fl ≥ 0.4 and Fl ≤ 0.6, we considered them as 'pseudo-heterozygous' variants, since such positions show about half of calls as 'variant' and about half of them as ‘non-variant’. Sites exhibiting Fl values out of ranges indicated above ($Fl \geq 0.1$ and $Fl < 0.4$ or $Fl > 0.6$ and $Fl < 0.9$), they have been considered as “variants to be validated”, in order to decide whether discarding them or be assigned, after validation, as ‘pseudo-heterozygous’ or ‘pseudo-homozygous’. These decisions were made also considering the phylogenetic context: variants shared among samples belonging to the same haplogroup are more likely to be true calls.

The final set of ‘pseudo-diploid’ variants which passed the filtering criteria were then manually checked in the alignment .bam file of samples. In the final decision, we considered several criteria such as the phylogenetic context, the depth and the quality of the region, the proximity of repetitive elements and the presence of the same variant position with suboptimal parameters in other subjects. We also retained the clustered variants, since the presence of clusters of mutation, occurring at closely spaced positions, may be indicative of the same mutational event, such as gene conversion.

Filtering of variants within palindrome spacer

Due to the haploid features of the spacer and its general high-quality reads, the filtering of variants found within the spacer is slightly different from that adopted for the arms.

We retained SNPs showing QUAL value ≥ 100, then we adopted the invariant following steps already used for duplicated regions:

- If $DP \geq 2$ and $DP_{ALT} \leq 2$: the variant is discarded;
- If $DP_{ALT}/DP_{REF} < 0.1$: the variant is discarded.

The list of variants was refined by calculating the Fl parameter: if the variants show Fl value < 0.3, we directly rejected them; if the variants exhibit Fl value > 0.8, we directly considered them as true.

In case of variant positions showing intermediate values of Fl (Fl ≥ 0.3 and Fl < 0.8), they underwent experimental validation.

Also in the case of spacers, the list of mutations which passed the filtering criteria was manually checked in the .bam files and analysed in the phylogenetic context. In this analysis we discarded mutations occurring at closely spaced positions (less than 20 bp) in single samples, being indicative of the same recombinative event.

Validation of variants through Sanger sequencing

We validated the genetic status of the variant positions showing intermediate parameters by means of PCRs and Sanger sequencing procedure.

All markers have been amplified following a standard protocol of touchdown PCR. The amplification reaction was performed starting from 50/100 ng of genomic DNA. The 20-mer primers selected for both amplification and sequencing have been designed referring to the GRCh37/hg19 human genome sequence and using Primer3 v. 0.4.0. software. Primers have been designed to specifically amplify the Y chromosome and discard the interspersed repeated elements. The specific Y chromosome amplification was confirmed by an in silico PCR with the UCSC Genome Browser tool, which returned 2 amplicons for palindrome arms and 1 for the spacer.

The purification of the PCR products and the sequencing reaction were carried out at Eurofins srl in Milan (http://www.eurofins.it) or at Bio-Fab Research srl in Rome (http://www.biofabresearch.it). Fluorescent sequencing reactions were performed and run on an automatic Applied Biosystems 3730xl DNA Analyzer using 20-mer internal oligonucleotides as sequencing primers. The sequences obtained were aligned and compared with Sequencher v. 4.8 (Gene Codes Corporation) in order to establish the allelic variants.

Analysis of the Y-Y gene conversion

Detection of PSVs and Y-Y gene conversion events

Gene conversion changes the state of a ‘pseudo-diploid’ genotype from heterozygous to homozygous. So, the detection of a gene conversion event strongly depends on the possibility to observe PSVs within the examined sequences, which designate ‘pseudo-heterozygous’ states. Generally, PSVs have been generated by a single mutational event on the proximal or on the distal arm of the palindrome. Thus, the possibility to find a gene conversion event does not depend on the arm where the mutation occurred. The minimum number of mutations (generating new PSVs) and gene conversion events is given by mapping each event within the phylogeny, according to the following criteria:

- When we observed a single chromosome showing a PSV, we considered it as the result of a single mutational event occurring on a palindrome arm of that chromosome. The observation of a phylogenetic cluster of chromosomes showing the same PSV indicates that the mutational event generating such PSV occurred at the branch joining all the interested chromosomes (Figure 22A). On the contrary, the same PSV shared between ≥ 2 phylogenetically unrelated chromosomes has been considered as generated by different mutational events occurred at different branches. We inferred the ancestral/derived state of PSVs according to their phylogenetic distribution;

- The observation of ‘pseudo-homozygous’ chromosomes descending from the branch where the PSV arose, is indicative that a gene conversion event (Figure 22B) (or more than one - Figure

22C) has occurred. In order to investigate the direction of the conversion events (ancestral to derived or vice versa), we used the ancestral/derived state information of the PSV (Figure 22B and C);

- About PSVs that arose before the human Y chromosome radiation (branch 1 - Figure 22C), due to the lack of information from deeper nodes of the phylogeny, we inferred the ancestral state of the variant and the direction of the gene conversion event referring to the orthologous base on the chimpanzee sequence (Clint_PTRv2/panTro6);

- Within a single PSV, the observation in the phylogeny of exclusively ‘pseudo-homozygous’ chromosomes showing different genotypes (Figure 22D) suggests that a mutational event generating a PSV and a subsequent gene conversion towards the derived state have occurred on the same branch of the phylogeny, in a close time frame.

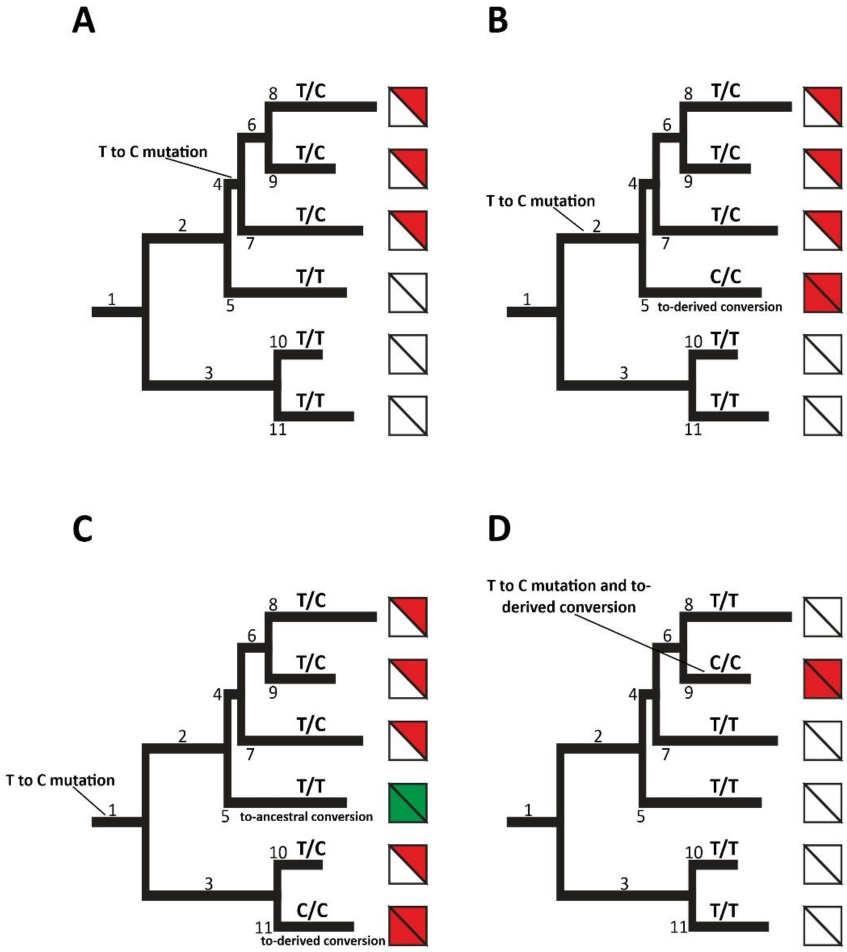

Figure 22: Identification of mutational events generating PSVs and of gene conversion events across the phylogeny. 'Pseudo-heterozygous' chromosomes are indicated with red triangles, 'pseudo-homozygous' non-mutated chromosomes are shown with white squares, whereas 'pseudo-homozygous' ancestral and derived converted chromosomes are represented by green and red squares, respectively. Branch nomenclature is reported within the tree and on terminal branches the 'pseudo-allelic' states of each chromosome is indicated. According to the most parsimonious explanation for the diversity pattern observed: **A)** A T to C mutation occurs along branch 4 creating a T/C PSV in 3 chromosomes; **B)** A T to C mutation occurs along branch 2 and a gene

conversion towards the derived state occurs on branch 5; **C)** T to C mutation occurring at the root of the tree introducing a PSV in all the descendant chromosomes, then two independent gene conversion events occur on branch 5 (ancestral) and 11 (derived); **D)** Mutational event and subsequent to-derived gene conversion, both occurring on branch 9 of the phylogeny.

Y-Y gene conversion rate estimate

Thanks to the estimate of the mutation number characterising each branch of the phylogeny (Figure 12) and of the average elapsed time for each mutational event in the X-degenerate region (406.6 years), we were able to estimate the time of each branch of the tree and to calculate a palindrome-specific gene conversion rate (*c*), according to the following equation:

$$c = \frac{\sum_{i=1}^{n} Ci}{\sum_{i=1}^{n} ti}$$

Where *C* is the number of the independent gene conversion events observed along the phylogeny which occurred within the i*th* PSV and *n* is the number of PSVs identified in each palindrome; *ti* is the time of existence of a single PSV within the phylogeny, calculated as the sum of the times of all branches in which the PSV is present, and corresponds to the time frame when is possible to observe a gene conversion event within the i*th* PSV, in years. According to this, our estimate consists in an “observed gene conversion rate”.

Through this approach, we performed two estimates of the gene conversion rate by calculating a minimum and maximum time of conversion. In the first estimate we included the time of the branch where the PSV arose and the time of the branch(es) where the gene conversion event(s) occurred, considering that the mutation generating such PSV may have occurred at the beginning time of the branch, whereas the gene conversion(s) converting such PSV may have occurred at the end time of the converted branch(es). In this way, we obtained a maximum conversion time (and a corresponding minimum conversion rate).

In the second estimate both these two were excluded, considering that each PSV may have arisen at the end time of the branch where it occurs and the subsequent gene conversion(s) may have occurred at the beginning time of the converted branch(es), thus presuming they took place really close in time. This returned a minimum conversion time (and a maximum gene conversion rate). As a consequence, we calculated the whole time of conversion as the average value of the two estimated times.

For instance, PSV represented in Figure 22A contributes with 0 gene conversion events and with a time corresponding to the sum of the time of branches 4, 6, 7, 8 and 9 to a maximum conversion time estimate. Then, for the same PSV, we estimated a minimum conversion time by excluding the time of branch 4, where the PSV arose. Accordingly, PSV shown in Figure 22C contributes with 2 gene conversion events, and with a time equal to the sum of the times of all branches of the tree (from branch 1 to branch 11) to generate a maximum conversion time. Similarly, we obtained a minimum conversion time for this PSV by excluding the time of branches 1, where the PSV arose, and of the converted branches 5 and 11. About PSV indicated in Figure 22D, we observe a single mutation and gene conversion event both occurring on branch 9, thus it contributes with 1 gene conversion event and with the time of branch 9 to the maximum conversion time estimate. On the contrary, a time = 0 for this PSV has been considered for the minimum conversion time estimate.

Estimate of the palindrome-specific mutation rate

We estimated the palindrome-specific mutation rate (*m*) within the arms as follows:

$$m = \frac{N}{t \times L}$$

We calculated the ratio between the total independent mutational events found across the phylogeny (*N*) and the whole time of phylogeny connecting the 161 Y chromosomes (*t*). The latter has been calculated as the total number of mutations found within the X-degenerate region × 406.6 years (time between two consecutive mutations). We estimated a per base mutation rate dividing by the length (in bp) of sequences under study (*L*).

The same method has been adopted to estimate the mutation rate within each palindrome spacer, calculated as the ratio between the total number of independent mutations found in the spacer and the whole time of the phylogeny. We obtained a per base rate dividing by the length (in bp) of the spacer.

Calculation of the Recurrence Index

In order to assess the extent of recurrence of mutations in palindromic regions, we calculated the Recurrence Index (*RI*%), that expresses the percentage of mutations that occur in positions that already had 1 mutation. It is calculated as follows:

$$RI(\%) = \frac{Me - Vp}{Me} \times 100$$

Where *Me* is the total number of mutational events, considering any event (mutation, Y-Y gene conversion and X-Y gene conversion) observed along the phylogeny that introduces a nucleotide change in the DNA sequence. *Vp* is the total number of variant positions observed along both arms. It is calculated as the sum of all variant bases found within all PSVs, considering that within each PSV, the number of variant positions can vary between

0 (in case of a monomorphic ancestral PSV) and 2 (maximum number of paralogous sites per PSV that can change).

Analysis of the X-to-Y gene conversion

We evaluated the extent of the X-to-Y gene conversion by performing four pair-wise alignments of the human palindrome P8 with each of the 4 gametologous sequences on the X chromosome, containing a copy of the VCX gene, respectively. To this aim, we searched for X-Y gametologous sequences by means of BLAT tool of the UCSC Genome Browser, then we used LAGAN program (Brudno et al. 2003) provided by VISTAtools to obtain the alignments.

The consequence of a gene conversion event from the X chromosome to a single arm of palindrome P8 is to increase the number of differences between arms, creating new PSVs or reintroducing them after a Y-Y conversion event. Thus, for each alignment, we investigated all possible Y-Y PSVs that may have been introduced by one of the donor X-linked sequences. More specifically, we searched for the derived state of P8 SNPs with the derived allele corresponding to the gametologous base on the X chromosome.

After this analysis, we reinterpreted the mutational patterns that led to the formation of Y-Y PSVs also taking into account the gene conversion from the X chromosome, and justifying the observed diversity with the least number of mutational steps.

<u>Identification of human-chimpanzee C-sites</u>

In order to identify inter-species historical gene conversion events within the X-Y gametologous region analysed, we used BLAT tool to select the four X-linked sequences and P8 palindrome from chimpanzee genome, known to be highly similar to human orthologs. Then, for each X-linked sequence, we performed a four-way alignment of sex chromosomes from both human and chimpanzee (Figure 23), by means of ClustalW tool.

Within each of the four alignments, we investigated the distribution of conversion sites (C-sites) using the same approach of Trombetta et al. (2014), describing molecular mechanisms for the formation of variable sites (Figure 23). More specifically, starting from an invariant site among species, a S-site may arise if a mutation occurs within a single sex chromosome after human-chimpanzee divergence (Figure 23), whereas a N-site may be generated by a mutation occurring before the speciation, being present in both X (or Y) chromosomes of the two species (Figure 23). If a gene conversion event involves a N-site, it generates a S-site independently of direction, whereas a conversion of a S-nucleotide may generate a C-site (Figure 23) or an invariant site depending on the direction of the gene conversion itself.

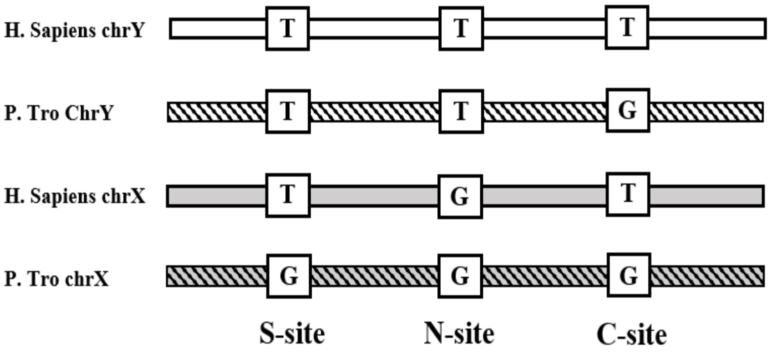

Figure 23: Possible variant sites within a four-way alignment of the orthologous and gametologous sequences from human-chimpanzee sex chromosomes. Different types of sites are shown: C-site or conversion site; N-site or non-conversion site; S-site or singleton (adapted from Trombetta et al. 2014).

X-to-Y gene conversion rate estimate

To estimate the rate of the X-to-Y gene conversion ($c_{x\text{-}y}$) within the gametologous region analysed, we used a slightly modified version of the method described in Cruciani et al. (2010b). We considered the length of the converted tracts and we divided by the time which span the tree connecting the 161 Y chromosomes. The equation for the rate is the following:

$$c_{x\text{-}y} = \frac{1}{Lt}\sum_{i=1}^{n} li$$

where $c_{x\text{-}y}$ is the estimated rate of gene conversion, *n* is the number of observed gene conversion events, *li* is the length in bp of the *i*th gene conversion event, *L* is the length (in bp) of the region under study, and *t* is the time of the tree, in years.

REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Adams SM et al. (2006). The case of the unreliable SNP: Recurrent back-mutation of Y-chromosomal marker P25 through gene conversion. Forensic Sci Int 159:14–20.

Ali S and Hasnain SE (2003). Genomics of the human Y-chromosome: 1. Association with male infertility. Gene 321:25–37.

Bagnall R et al. (2005). Gene conversion and evolution of Xq28 duplicons involved in recurring inversions causing severe hemophilia A. Genome Res 15:214–223.

Bailey JA and Eichler EE (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552–564.

Bailey JA et al. (2002). Recent Segmental Duplications in the Human Genome. Science 297:1003.

Balaresque P et al. (2014). Gene Conversion Violates the Stepwise Mutation Model for Microsatellites in Y-Chromosomal Palindromic Repeats. Hum Mutat 35:609–617.

Bandelt HJ et al. (1999). Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48.

Barbieri C et al. (2016). Refining the Y chromosome phylogeny with southern African sequences. Hum Genet 135:541–553.

Ben Khelifa H et al. (2013). Xp22.3 interstitial deletion: A recognizable chromosomal abnormality encompassing VCX3A and STS genes in a patient with X-linked ichthyosis and mental retardation. Gene 527:578–583.

Betrán E et al. (2012). Why Chromosome Palindromes? Int J Evol Biol 2012:207958.

Bhowmick BK et al. (2006). Comparative analysis of human masculinity. Genet Mol Res 5:696-712.

Bhowmick BK et al. (2007). The origin and evolution of human ampliconic gene families and ampliconic structure. Genome Res 17:441–450.

Bischof JM et al. (2006). Genome-wide identification of pseudogenes capable of disease-causing gene conversion. Hum Mutat 27:545–552.

Blanco P et al. (2000). Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J Med Genet 37:752.

Bosch E et al. (2004). Dynamics of a human interparalog gene conversion hotspot. Genome Res 14:835–844.

Brudno M et al. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731.

Carvalho CMB and Lupski JR (2016). Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17:224–238.

Casola C et al. (2012). Interlocus gene conversion events introduce deleterious mutations into at least 1% of human genes associated with inherited disease. Genome Res 22:429–435.

Cechova M et al. (2020). Dynamic evolution of great ape Y chromosomes. Proc Natl Acad Sci U S A.

Chen JM et al. (2007). Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8:762–775.

Chiaroni J et al. (2009). Y chromosome diversity, human expansion, drift, and cultural evolution. Proc Natl Acad Sci U S A 106:20174–20179.

Chiaroni J et al. (2010). The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations. Eur J Hum Genet 18:348–353.

Conrad B and Antonarakis SE. (2007). Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. Annu Rev Genomics Hum Genet 8:17–35.

Costa P et al. (2008). Identification of new breakpoints in AZFb and AZFc. Mol Hum Reprod 14:251–258.

Cotter DJ et al. (2016). Genetic Diversity on the Human X Chromosome Does Not Support a Strict Pseudoautosomal Boundary. Genetics 203:485–492.

Cruciani F et al. (2002). A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes. Am J Hum Genet 70:1197–1214.

Cruciani F et al. (2004). Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am J Hum Genet 74:1014–1022.

Cruciani F et al. (2006). Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. Hum Mutat 27:831–832.

Cruciani F et al. (2007). Tracing Past Human Male Movements in Northern/Eastern Africa and Western Eurasia: New Clues from Y-Chromosomal Haplogroups E-M78 and J-M12. Mol Biol Evol 24:1300–1311.

Cruciani F et al. (2008). Recurrent mutation in SNPs within Y chromosome E3b (E-M215) haplogroup: A rebuttal. Am J Hum Biol 20:614–616.

Cruciani F et al. (2010a). Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. Eur J Hum Genet 18:800–807.

Cruciani F et al. (2010b). About the X-to-Y Gene Conversion Rate. Am J Hum Genet 86:495–497.

Cruciani F et al. (2011). A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. Am J Hum Genet 88:814–818.

D’Atanasio E et al. (2018). The peopling of the last Green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages. Genome Biol 19:20–20.

Davis JK et al. (2010). A W-linked palindrome and gene conversion in New World sparrows and blackbirds. Chromosome Res 18:543–553.

De Marco P et al. (2000). Folate pathway gene alterations in patients with neural tube defects. Am J Med Genet 95:216–223.

Deng H et al. (2018). Histone H3.3K27M Mobilizes Multiple Cancer/Testis (CT) Antigens in Pediatric Glioma. Mol Cancer Res 16:623.

Dennis MY and Eichler EE (2016). Human adaptation and evolution by segmental duplication. Curr Opin Genet Dev 41:44–52.

Dhanoa JK et al. (2016). Y-chromosomal genes affecting male fertility: A review. Vet World 9:783–791.

Di Rienzo A et al. (1994). Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A 91:3166–3170.

Drummond AJ and Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214.

Duret L and Galtier N (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annu Rev Genom Hum Genet 10:285–311.

Eichler EE (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet 17:661–669.

Estivill X and Armengol L (2007). Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet 3:1787–1799.

Ezawa K et al. (2006). Genome-Wide Search of Gene Conversions in Duplicated Genes of Mouse and Rat. Mol Biol Evol 23:927–940.

Feuk L et al. (2006). Structural variation in the human genome. Nat Rev Gen 7:85–97.

Finocchio A et al. (2018). A finely resolved phylogeny of Y chromosome Hg J illuminates the processes of Phoenician and Greek colonizations in the Mediterranean. Sci Rep 8:7465–7465.

Forster P et al. (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935–945.

Francalacci P et al. (2013). Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341:565–569.

Freije D et al. (1992). Identification of a second pseudoautosomal region near the Xq and Yq telomeres. Science 258:1784.

Friães A et al. (2006). CYP21A2 mutations in Portuguese patients with congenital adrenal hyperplasia: Identification of two novel mutations and characterization of four different partial gene conversions. Mol Genet Metab 88:58–65.

Fu Q et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449.

Galtier N (2003). Gene conversion drives GC content evolution in mammalian histones. Trends Genet 19:65–68.

Galtier N et al. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics 159:907.

Geraldes A et al. (2010). Extensive Gene Conversion Drives the Concerted Evolution of Paralogous Copies of the SRY Gene in European Rabbits. Mol Biol Evol 27:2437–2440.

Graves JAM (2006). Sex Chromosome Specialization and Degeneration in Mammals. Cell 124:901–914.

Grugni V et al. (2019). Analysis of the human Y-chromosome haplogroup Q characterizes ancient population movements in Eurasia and the Americas. BMC Biol 17:3.

Gusmão L et al. (2005). Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26:520–528.

Haber JE et al. (2004). Repairing a double-strand chromosome break by homologous recombination: revisiting Robin Holliday’s model. Philos Trans R Soc Lond B Biol Sci 359:79–86.

Hallast P and Jobling MA (2017). The Y chromosomes of the great apes. Hum Genet 136:511–528.

Hallast P (2005). Segmental duplications and gene conversion: Human luteinizing hormone/chorionic gonadotropin gene cluster. Genome Res 15:1535–1546.

Hallast P et al. (2013). Recombination Dynamics of a Human Y-Chromosomal Palindrome: Rapid GC-Biased Gene Conversion, Multi-kilobase Conversion Tracts, and Rare Inversions. PLoS Genet 9:e1003666.

Hallast P et al. (2015). The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol 32:661–673.

Hammer MF and Zegura SL (2002). The Human Y Chromosome Haplogroup Tree: Nomenclature and Phylogeography of Its Major Divisions. Annu Rev Anthropol 31:303–321.

Hammer MF (1995). A recent common ancestry for human Y chromosomes. Nature 378:376–378.

Hammer MF et al. (2000). Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci U S A 97:6769–6774.

Harpak A et al. (2017). Frequent non-allelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. Proc Natl Acad Sci U S A 114:12779–12784.

Helgason A et al. (2015). The Y-chromosome point mutation rate in humans. Nat Genet 47:453–457.

Hosomi N et al. (2007). Deletion of distal promoter of VCXA in a patient with X-linked ichthyosis associated with borderline mental retardation. J Dermatol Sci 45:31–36.

Hughes JF et al. (2005). Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. Nature 437:100–103.

Hughes JF et al. (2010). Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature 463:536–539.

Hughes JF et al. (2012). Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature 483:82–86.

Innan H and Kondrashov F (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108.

International Human Genome Sequencing Consortium et al. (2001). Initial sequencing and analysis of the human genome. Nature 409:860–921.

Ira G et al. (2006). Conservative inheritance of newly synthesized DNA in double-strand break-induced gene conversion. Mol Cell Biol 26:9424–9429.

Jain M et al. (2018). Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol 36:321–323.

Jiang Z et al. (2007). Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet 39(11):1361–1368.

Jobling MA and Tyler-Smith C (2003). The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598–612.

Jobling MA and Tyler-Smith C (2017). Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet 18:485–497.

Jobling MA (2008). Copy number variation on the human Y chromosome. Cytogenet Genome Res 123:253–262.

Jobling MA et al. (2004). Human evolutionary genetics. Garland Science, New York.

Jobling MA et al. (2013). Human evolutionary genetics (second edition). Garland Science: New York.

Jones ER et al. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat Commun 6:8912.

Karafet T et al. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18:830–838.

Karmin M et al. (2015). A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res 25:459–466.

Kayser M et al. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66:1580–1588.

Kayser M et al. (2004). A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet 74:1183–1197.

Kayser M et al. (2006). Melanesian and Asian Origins of Polynesians: mtDNA and Y Chromosome Gradients Across the Pacific. Mol Biol Evol 23:2234–2244.

Kehrer-Sawatzki H and Cooper DN (2007). Structural divergence between the human and chimpanzee genomes. Hum Genet 120:759–778.

Kong A et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature 488:471–475.

Kudla G et al. (2004). Gene Conversion and GC-Content Evolution in Mammalian Hsp70. Mol Biol Evol 21:1438–1444.

Kumar S et al. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870–1874.

Kuroda-Kawaguchi T et al. (2001). The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet 29:279–286.

Kuroki Y et al. (2006). Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. Nat Genet 38:158–167.

Lahn BT and Page DC (1999). Four Evolutionary Strata on the Human X Chromosome. Science 286:964.

Lahn BT and Page DC (2000). A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. Hum Mol Genet 9:311–319.

Lahn BT et al. (2001). The human Y chromosome, in the light of evolution. Nat Rev Genet 2:207–216.

Lange et al. (2009). Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. Cell 138:855–69.

Lartillot N (2013). Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. Mol Biol Evol 30:489–502.

Lazaridis I et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413.

Levinson G and Gutman GA (1987). High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. Nucleic Acids Res 15:5323–5338.

Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li H (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

Li H et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Lupski JR (2004). Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. Genome Biol 5:242.

Mancera E et al. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479–485.

Marais G (2003). Biased gene conversion: implications for genome and sex evolution. Trends Genet 19:330–338.

Massaia A and Xue Y (2017). Human Y chromosome copy number variation in the next generation sequencing era and beyond. Hum Genet 136:591–603.

Mendez FL et al. (2013). An African American Paternal Lineage Adds an Extremely Ancient Root to the Human Y Chromosome Phylogenetic Tree. Am J Hum Genet 92:454–459.

Mendez FL et al. (2016). The Divergence of Neandertal and Modern Human Y Chromosomes. Am J Hum Genet 98:728–734.

Mohyuddin A et al. (2006). Detection of novel Y SNPs provides further insights into Y chromosomal variation in Pakistan. J Hum Genet 51:375–378.

Morton NE (1991). Parameters of the human genome. Proc Natl Acad Sci U S A 88:7474–7476.

Myres NM et al. (2011). A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet 19:95–101.

Nakashima E et al. (2004). Novel SBDS mutations caused by gene conversion in Japanese patients with Shwachman-Diamond syndrome. Hum Genet 114:345–348.

Newman TL et al. (2005). A genome-wide survey of structural variation between human and chimpanzee. Genome Res 15:1344–1356.

Page DC et al. (1984). Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. Nature 311:119–123.

Page DC et al. (1987). Linkage, physical mapping, and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. Genomics 1:243–256.

Papadakis MN and Patrinos GP (1999). Contribution of gene conversion in the evolution of the human β-like globin gene family. Hum Genet 104:117–125.

Pearson CE et al. (1996). Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. J Cell Biochem 63:1–22.

Pentao L et al. (1992). Charcot–Marie–Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. Nat Genet 2:292–300.

Poznik GD et al. (2013). Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341:562–565.

Poznik GD et al. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet 48:593–599.

Reiter LT et al. (1998). Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. Am J Hum Genet 62:1023–1033.

Rice WR (1987). Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. Genetics 116:161–167.

Rivera Franco N et al. (2020). Identifying new lineages in the Y chromosome of Colombian Amazon indigenous populations. Am J Phys Anthropol 172:165–175.

Ross MT et al. (2005). The DNA sequence of the human X chromosome. Nature 434:325–337.

Rosser ZH et al. (2009). Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. Am J Hum Genet 85:130–134.

Rozen S et al. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873–876.

Sachidanandam R et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933.

Saillard J et al. (2000). mtDNA Variation among Greenland Eskimos: The Edge of the Beringian Expansion. Am J Hum Genet 67:718–726.

Sainz J et al. (2006). Segmental duplication density decrease with distance to human-mouse breaks of synteny. Eur J Hum Genet 14:216–221.

Samonte RV and Eichler EE (2002). Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3:65–72.

Schildkraut E et al. (2005). Gene conversion and deletion frequencies during double-strand break repair in human cells are controlled by the distance between direct repeats. Nucleic Acids Res 33:1574–1580.

Scozzari R et al. (2012). Molecular Dissection of the Basal Clades in the Human Y Chromosome Phylogenetic Tree. PLoS One 7:e49170.

Scozzari R et al. (2014). An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res 24:535–544.

Sharp AJ et al. (2005). Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88.

Shen P et al. (2000). Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci U S A 97:7354–7359.

Shen P et al. (2004). Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-Chromosome and mitochondrial DNA sequence Variation. Hum Mutat 24:248–260.

Shi W et al. (2018). Copy number variation arising from gene conversion on the human Y chromosome. Hum Genet 137:73–83.

Shi W et al. (2019). Birth, expansion, and death of VCY-containing palindromes on the human Y chromosome. Genome Biol 20:207.

Shrivastav M et al. (2008). Regulation of DNA double-strand break repair pathway choice. Cell Res 18:134–147.

Skaletsky H et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825–837.

Skinner BM et al. (2016). The pig X and Y Chromosomes: structure, sequence, and evolution. Genome Res 26:130–139.

Skov L et al. (2017). Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. PLoS Genet 13:e1006834.

Soh YQS et al. (2014). Sequencing the Mouse Y Chromosome Reveals Convergent Gene Acquisition and Amplification on Both Sex Chromosomes. Cell 159:800–813.

Stankiewicz P and Lupski JR (2002). Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev 12:312–319.

Stankiewicz P et al. (2004). Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res 14:2209–2220.

Stone AC et al. (2002). High levels of Y-chromosome nucleotide diversity in the genus Pan. Proc Natl Acad Sci U S A 99:43–48.

Sun C et al. (1999). An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y. Nat Genet 23:429–432.

Surdhar GK et al. (2001). Homozygous gene conversion in von Willebrand factor gene as a cause of type 3 von Willebrand disease and predisposition to inhibitor development. Blood 98:248–250.

Swanepoel CM et al. (2020). Large X-Linked Palindromes Undergo Arm-to-Arm Gene Conversion across Mus Lineages. Mol Biol Evol 37:1979–1985.

Szostak JW et al. (1983). The double-strand-break repair model for recombination. Cell 33:25–35.

Taguchi A et al. (2014). A Search for Novel Cancer/Testis Antigens in Lung Cancer Identifies VCX/Y Genes, Expanding the Repertoire of Potential Immunotherapeutic Targets. Cancer Res 74:4694.

Tamura K et al. (2011). MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 28:2731–2739.

Teitz LS et al. (2018). Selection Has Countered High Mutability to Preserve the Ancestral Copy Number of Y Chromosome Amplicons in Diverse Human Lineages. Am J Hum Genet 103:261–275.

Tomaszkiewicz M et al. (2016). A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y. Genome Res 26:530–540.

Trombetta B and Cruciani F (2017). Y chromosome palindromes and gene conversion. Hum Genet 136:605–619.

Trombetta B et al. (2010). Footprints of X-to-Y Gene Conversion in Recent Human Evolution. Mol Biol Evol 27:714–725.

Trombetta B et al. (2011). A New Topology of the Human Y Chromosome Haplogroup E1b1 (E-P2) Revealed through the Use of Newly Characterized Binary Polymorphisms. PLoS ONE 6:e16073.

Trombetta B et al. (2014). Inter- and Intraspecies Phylogenetic Analyses Reveal Extensive X–Y Gene Conversion in the Evolution of Gametologous Sequences of Human Sex Chromosomes. Mol Biol Evol 31:2108–2123.

Trombetta B et al. (2015a). Phylogeographic Refinement and Large Scale Genotyping of Human Y Chromosome Haplogroup E Provide New Insights into the Dispersal of Early Pastoralists in the African Continent. Genome Biol Evol 7:1940–1950.

Trombetta B et al. (2015b). Regional Differences in the Accumulation of SNPs on the Male-Specific Portion of the Human Y Chromosome Replicate Autosomal Patterns: Implications for Genetic Dating. PLoS One 10:e0134646–e0134646.

Trombetta B et al. (2016). Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. Sci Rep 6:28710–28710.

Underhill PA and Kivisild T (2007). Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. Annu Rev Genet 41:539–564.

Underhill PA et al. (1996). A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci U S A 93:196–200.

Underhill PA et al. (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7:996–1005.

Underhill PA et al. (2000). Y chromosome sequence variation and the history of human populations. Nat. Genet. 26: 358-361.

Underhill PA et al. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet 65:43–62.

van den Bosch M et al. (2002). DNA double-strand break repair by homologous recombination. Biol Chem 383:873–892.

Van Esch H et al. (2005). Deletion of VCX-A due to NAHR plays a major role in the occurrence of mental retardation in patients with X-linked ichthyosis. Hum Mol Genet 14:1795–1803.

Vanita Sarhadi V et al. (2001). A unique form of autosomal dominant cataract explained by gene conversion between beta-crystallin B2 and its pseudogene. J Med Genet 38:392–396.

Ventura M et al. (2011). Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. Genome research 21(10):1640–1649.

Verrelli BC and Tishkoff SA (2004). Signatures of selection and gene conversion associated with human color vision variation. Am J Hum Genet 75:363–375.

Wang S et al. (2018). De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. Cell Rep 24:3441–3454.

Warburton PE et al. (2004). Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res 14:1861–1869.

Weber JL and Wong C (1993). Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128.

Wei W et al. (2013). A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. Forensic Sci Int Genet 7:568–572.

Whitfield LS et al. (1995). Sequence variation of the human Y chromosome. Nature 378:379–380.

Woelk CH et al. (2007). Evolution of the interferon alpha gene family in eutherian mammals. Gene 397:38–50.

Woodward KJ et al. (2019). Atypical nested 22q11.2 duplications between LCR22B and LCR22D are associated with neurodevelopmental phenotypes including autism spectrum disorder with incomplete penetrance. Mol Genet Genomic Med 7:e00507.

Xue Y et al. (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol 19:1453–1457.

Y Chromosome Consortium (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12:339–348.

Yuan B et al. (2015). Comparative Genomic Analyses of the Human NPHP1 Locus Reveal Complex Genomic Architecture and Its Regional Evolution in Primates. PLoS Genet 11:e1005686–e1005686.

Zangenberg et al. (1995). New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. Nat Genet. 10:407–414.

Zhang L et al. (2004). Patterns of Segmental Duplication in the Human Genome. Mol Biol Evol 22:135–141.

Zhou R et al. (2020). A willow sex chromosome reveals convergent evolution of complex palindromic repeats. Genome Biol 21:38.

Zou SW et al. (2003). Expression and localization of VCX/Y proteins and their possible involvement in regulation of ribosome assembly during spermatogenesis. Cell Res 13:171–177.

Web resources

ISOGG	http://isogg.org/
MEGA:	http://www.megasoftware.net
Network:	http://www.fluxus-engineering.com
Primer3:	http://bioinfo.ut.ee/primer3-0.4.0/
UCSC Genome browser:	http://genome.ucsc.edu/

ADDITIONAL FILES

Additional Figure 1. ‘Pseudo-diploid’ positions identified in palindrome P6 arms. To the left, it is reported the Y chromosome tree showing the phylogenetic relationships among the 157 samples. SNP names are given at the top. Each square is divided into two triangles, representing the paralogous sites in the two arms of the palindrome.
https://drive.google.com/drive/folders/1GTnhTR_p1qQ2zUCi9AM_iuO06zGs076E?usp=sharing

Additional Figure 2. ‘Pseudo-diploid’ positions identified in palindrome P8 arms. To the left, it is reported the Y chromosome tree showing the phylogenetic relationships among the 157 samples. SNP names are given at the top. Each square is divided into two triangles, representing the paralogous sites in the two arms of the palindrome.
https://drive.google.com/file/d/17YWRVdl9H8mBX1Q5hPEI9yL8LrC_xksx/view?usp=sharing

Additional File 1. List of mutations and gene conversion events found in P6 palindrome arms.
https://drive.google.com/file/d/1xDtYReofw-furYJdl1gxobExcfZBF5QT/view?usp=sharing

Additional File 2. List of mutations and gene conversion events found in P7 palindrome arms.
https://drive.google.com/file/d/1H5HiLQpXdTpilVjj_X0T7fo3iKfWRHlD/view?usp=sharing

Additional File 3. List of mutations and gene conversion events found in P8 palindrome arms.
https://drive.google.com/file/d/1SHOusdl4SKje0cS6537-zW0s6vEsSuVM/view?usp=sharing

LIST OF PUBLICATIONS

1. D'Atanasio E, Trionfetti F, **Bonito M**, Sellitto D, Coppa A, Berti A, Trombetta B, Cruciani F (2020). Y Haplogroup Diversity of the Dominican Republic: Reconstructing the Effect of the European Colonization and the Trans-Atlantic Slave Trades. *Genome Biology and Evolution*, 12:1579–1590.

2. Della Rocca C, Cannone F, D'Atanasio E, **Bonito M**, Anagnostou P, Russo G, Barni F, Alladio E, Destro-Bisol G, Trombetta B, Berti A, Cruciani F (2020). Ethnic fragmentation and degree of urbanization strongly affect the discrimination power of Y-STR haplotypes in central Sahel. *Forensic Science International Genetics*, 49:102374.

3. D'Atanasio E, Iacovacci G, Pistillo R, **Bonito M**, Dugoujon JM, Moral P, El-Chennawi F, Melhaoui M, Baali A, Cherkaoui M, Sellitto D, Trombetta B, Berti A, Cruciani F (2019). Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa. *Forensic Science International Genetics,* 38:185–194.

4. D’Atanasio E, **Bonito M**, Iacovacci G, Berti A, Trombetta B, Cruciani F (2019). Identification and molecular characterisation of an AMEL-X null allele due to an Alu insertion. *Forensic Science International Genetics*, 38:e1–e4.

5. D'Atanasio E, Trombetta B, **Bonito M**, Finocchio A, Di Vito G, Seghizzi M, Romano R, Russo G, Paganotti GM, Watson E, Coppa A, Anagnostou P, Dugoujon JM, Moral P, Sellitto D, Novelletto A, Cruciani F (2018). The peopling of the last green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages. *Genome Biology,* 19:1–15.